

Natural selection and demography in ancient human introgression

Von der Fakultät für Lebenswissenschaften
der Universität Leipzig

genehmigte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

Dr. rer. nat.

vorgelegt von

Magister Martin Petr

geboren am 27. September 1986 in Třinec (Tschechische Republik)

Dekan: Prof. Dr. Marc Schönwiesner

Gutachter: Prof. Dr. Svante Pääbo

Prof. Dr. Montgomery Slatkin

Tag der Verteidigung: 26. Februar 2021

BIBLIOGRAPHISCHE DARSTELLUNG

Martin Petr

Natural selection and demography in ancient human introgression

Fakultät für Lebenswissenschaften, Universität Leipzig

Kumulative Dissertation

192 Seiten, 128 Literaturangaben, 54 Abbildungen, 14 Tabellen

The ability to recover ancient DNA from skeletal material has completely transformed the field of evolutionary anthropology, making it possible to sequence the genomes of individuals who lived thousands of years ago. In addition to resolving the long-standing question of admixture between Neanderthals and modern humans and uncovering evidence of dramatic migration events throughout human history, ancient DNA has become an important resource for understanding many facets of natural selection, which is often challenging using present-day genetic variation alone.

Chapter 1 examines the dynamics of negative selection acting against Neanderthal ancestry in modern humans and establishes its limits over long evolutionary timescales. It shows that the previously reported monotonic decline in Neanderthal ancestry over the last fifty-thousand years, thought to have been a result of negative selection, is a statistical artifact caused by incorrect assumptions about modern human demographic history, in particular the gene flow between Africa and West Eurasia. Re-estimation of the Neanderthal ancestry proportions over time using a more robust statistic no longer infers a significant decline of Neanderthal ancestry, which is shown to be consistent with simulations of negative selection across a wide range of selection parameters.

Chapter 2 describes the first comprehensive analysis of the Y chromosomes of Neanderthals and Denisovans. Although Neanderthals and Denisovans form a sister group to modern humans at the autosomal level, Neanderthal Y chromosomes are more similar to modern human rather than Denisovan Y chromosomes. In fact, the Y chromosomes of late Neanderthals represent a lineage introgressed from an early modern human population. This introgression, which occurred hundreds of thousands of years ago, completely replaced the Y chromosomes of early Neanderthals, mirroring the observations made from mitochondrial DNA. Population genetic simulations of selection and introgression show that although a complete replacement of both mitochondrial DNA and Y chromosomes is unlikely under neutrality, higher deleterious burden of Neanderthals predicts a rapid replacement of both loci by their modern human counterparts.

Finally, *Chapter 3* presents an R package *admixr*, designed to facilitate programming of automated, fully reproducible population genetic analyses using ADMIXTOOLS, a suite of programs widely used in ancient DNA research.

This thesis is based on the following publications:

1. **Petr, M.**, Pääbo, S., Kelso, J. & Vernot, B. **Limits of long-term selection against Neandertal introgression.** *PNAS*, 116, 1639–1644 (2019).
2. **Petr M.**, Hajdinjak M., Fu Q., Essel E., Rougier H., Crevecoeur I., Semal P., Golovanova L. V., Doronichev V. B., Lalueza-Fox C., de la Rasilla M., Rosas A., Shunkov M. V., Kozlikin M. B., Derevianko A. P., Vernot B., Meyer M., Kelso J. **The evolutionary history of Neanderthal and Denisovan Y chromosomes.** *Science*, 369, 1653–1656 (2020).
3. **Petr, M.**, Vernot, B. & Kelso, J. ***admixr*—R package for reproducible analyses using ADMIXTOOLS.** *Bioinformatics*, 35, 3194–3195 (2019).

Contents

Summary	1
Zusammenfassung	13
Chapter 1	27
Chapter 2	61
Chapter 3	161
Bibliography	170
Curriculum Vitae	181
Declaration of Independent Work	186
Author Contribution Statements	188

Summary

Introduction

In 1856, an unusual skeleton was discovered in the Neander valley in today's Germany. This individual, later designated as the type specimen *Feldhofer 1*, as well as many others found during the decades after, belonged to a hominin group called Neanderthals (1). From a morphological perspective, these people carried characteristics undoubtedly similar to ours, yet they exhibited features as well as cultural practices distinct enough from modern humans¹ to fuel fierce debates for more than a hundred years that followed. Were Neanderthals a mere transition state on the way towards the modern human phenotype or did they represent an entirely new species of humans (2)?

As evidence mounted over decades, it became increasingly clear that Neanderthals and modern humans represent independently evolved and contemporaneous lineages (3, 4). Indeed, having occupied vast stretches of the Eurasian continent (5) after first emerging around 400 thousand years ago (kya) in Europe (6, 7), Neanderthals disappeared from the fossil record at around 40 kya (8), several thousand years after the arrival of the first modern humans into Eurasia (9–13). Thus, one of the most important topics in anthropology has become determining to which extent, if at all, Neanderthals interacted with modern humans during their millennia-long overlap in Eurasia. Have Neanderthals been completely replaced by modern humans without leaving a trace as modern humans migrated out of Africa to colonize the world (14)? Did the two groups of humans interbreed and have offspring? Could we detect any traces of Neanderthal genetic legacy in modern human genomes?

Morphological analyses of some of the earliest modern human fossils found in Europe found evidence of Neanderthal-derived morphological traits, suggesting that some interbreeding took place (9, 15–21), although the Neanderthal-modern human “hybrid” status of some of these individuals has been challenged (22–24). Furthermore, other evidence indicated the possibility of not only genetic but also cultural exchanges between Neanderthals and modern humans (25). Ultimately, however, to settle the question of whether Neanderthals contributed DNA to modern humans, and whether this genetic legacy still remains in people living today, required a genetic argument.

The first study of ancient DNA was published in 1984 and showed that DNA not only survived in a museum specimen of an extinct quagga almost 150 years after its death but that it could be successfully amplified by molecular cloning and sequenced (26). A year later, DNA extraction from Egyptian mummies demonstrated that the same could apply to museum specimens several thousand years older (27). However, the true revolution in the budding

¹ In this thesis, the phrase *modern human* is used as a shorter form of *anatomically modern human*. In this context, the word “modern” refers to a set of morphologic traits typical of present-day humans, a slightly unfortunate historical artifact that has become common in anthropological literature.

field of ancient DNA came with the development of the polymerase chain reaction (PCR) (28), which allowed the first DNA sequence of a Neanderthal to be determined (29).

This sequence, belonging to the *Feldhofer 1* type specimen, represented 379 bp of the mitochondrial DNA (mtDNA) control region, and its phylogenetic comparison to modern human mtDNAs showed that the Neanderthal mtDNA diverged prior to the diversification of present-day mtDNA lineages (29). Sequencing complete mtDNA genomes of many other Neanderthals over the following decade has always reached the same conclusion (5, 30–32) and, in the context of the wider anthropological debate, supported the hypothesis of the replacement of Neanderthals by modern humans without any mixing (29). However, a theoretical analysis of the first Neanderthal mtDNA study (29) cautioned that although the mtDNA phylogenetic relationships rejected models of random mating between Neanderthals and modern humans, there were other scenarios of gene flow which had to be considered (33). Among those, only the models of very large contributions of Neanderthal mtDNA to modern humans could be rejected given the limited statistical power from the available data and due to the narrow view of human evolutionary history provided by mtDNA alone (30, 33). In order to fully resolve the issue of Neanderthal gene flow into modern humans, information from the nuclear genome, which represents a composite of thousands of genealogies of any individual's ancestors, was needed.

The first analysis of the draft nuclear genome of a Neanderthal was published in 2010 (34). Finished less than ten years after the assembly of the first complete draft of the modern human genome (35), it represented a remarkable technical accomplishment requiring a large coordinated effort (36, 37). The comparison of the Neanderthal genome with the genomes of present-day humans showed that Neanderthals share more alleles with present-day non-Africans than with Africans, and that the majority of this signal must originate from Neanderthal gene flow into the ancestors of non-Africans (34). Later technical developments have allowed the sequencing of high-quality genomes of three additional Neanderthals, which led to the estimates of the proportion of Neanderthal ancestry in present-day non-Africans at around 2%, and to the inferred split time between Neanderthals and the ancestors of modern humans between 520 and 630 kya (38–40).

The definitive genetic evidence for introgression from Neanderthals into modern humans resolved one of the most debated issues in evolutionary anthropology (34). However, ancient DNA has continued to provide new insights into human evolutionary history. In 2010, DNA extracted from a distal manual phalanx unearthed in the Denisova cave in Siberia revealed an unknown type of the mitochondrial genome which diverged from a common ancestor with Neanderthal and modern human mtDNA sequences around 1 million years ago, raising the possibility that its carrier belonged to an entirely new hominin group (41). Indeed, sequencing of the nuclear genome of this individual assigned her to a yet unknown hominin

population that was a sister group to Neanderthals (38, 42, 43). Remarkably, this population, called the Denisovans, was found to have contributed up to 6% to the ancestry of people in present-day Papua New Guinea and Bougainville Islanders (42, 43). Later analysis of the geographical distribution of Denisovan ancestry in a larger sampling of Southeast Asian and Oceanian populations showed that the gene flow from Denisovans is likely to have occurred in Southeast Asia (44). As Denisovans have so far only been found in Siberia and on the Tibetan Plateau (45), this implies that they must have occupied an extraordinarily large geographical region, from Siberia to tropical Asia (44).

Having established that modern humans interbred with archaic humans as they migrated out of Africa (34, 42, 46), the focus of study of ancient human introgression shifted towards more nuanced questions. When and where did the interbreeding happen? How frequently did archaic and modern humans interact? Were there any functional consequences of introgressed archaic human DNA on modern human physiology?

Because of the equal affinity of the Neanderthal genome to the genomes of several Europeans and East Asians, the first draft Neanderthal genome study concluded that the gene flow likely happened somewhere in the Middle East, after the out-of-Africa migration but before the divergence of present-day non-African populations (34). This was corroborated by a later analysis of the distribution of lengths of Neanderthal haplotypes in present-day humans, which arrived at the time of introgression of about 55 thousand years ago (46). Intriguingly, later analyses of the geographic distribution of the proportion of Neanderthal ancestry in present-day genomes from around the globe (43, 47, 48) revealed a more complex picture of the history of Neanderthal admixture. Most importantly, East Asian populations appeared to carry approximately 20% more Neanderthal ancestry than Europeans (39, 43, 49–52).

Three major hypotheses have competed to explain the excess of Neanderthal ancestry in East Asians. The first suggested that the ancestors of East Asians interbred with Neanderthals a second time, after the separation from Europeans, which led to them carrying an additional fraction of Neanderthal ancestry (43). Alternatively, following the introgression into the ancestors of all non-Africans, Neanderthal ancestry in Europeans could have been “diluted” by admixture with a population which carried little to no Neanderthal ancestry, possibly some African group (43, 53). Finally, the larger amount of Neanderthal ancestry observed in East Asians was proposed to be a consequence of weaker selection against Neanderthal DNA in East Asians compared to Europeans due to the lower effective population size in the ancestors of East Asians (51). Population genetic modeling of the interaction of selection efficacy and genetic drift has argued against the selection model, and suggested that a second “pulse” of Neanderthal introgression into the ancestors of East Asians or the “dilution” scenario are more plausible (50, 54, 55).

To complicate matters even further, a recent study has shown that the excess of Neanderthal ancestry in East Asians today compared to Europeans might be, in fact, largely a statistical artifact (56). Specifically, most Neanderthal ancestry detection methods rely on the assumption that certain African groups do not carry any Neanderthal DNA at all, utilizing them as a “baseline” for detecting Neanderthal segments in other populations (50, 51, 56, 57). However, most African groups appear to carry some Neanderthal DNA resulting not from the Neanderthal introgression but from back migration from Eurasia into Africa (58, 59). Because this backflow is currently thought to have originated from a West Eurasian source, potentially even ancestors of Europeans themselves, this causes a bias against detecting Neanderthal ancestry in present-day Europeans (56, 59). Indeed, when using methods that do not make explicit assumptions about the complexities of modern human demographic history, the apparent excess of Neanderthal ancestry in East Asians largely disappears, reducing the need for a large role of dilution or multiple Neanderthal introgression events into the ancestors of present-day populations (56). My own work, described in detail in *Chapter 1*, has provided new evidence for gene flow from West Eurasia into Africa, highlighting the need for methods to estimate Neanderthal ancestry which are robust to these issues (59). Overall, the question of exactly how many “pulses” of Neanderthal (or Denisovan) ancestry have occurred during the time of overlap of archaic and modern humans in Eurasia remains unclear, but the evidence for multiple introgression events is growing stronger (12, 52, 60, 61).

In addition to detailing many aspects of past archaic and modern human encounters from a demographic perspective, ancient DNA has provided many insights into the functional consequences of Neanderthal and Denisovan DNA on modern human biology (62–64). This is partly thanks to large-scale databases of thousands of genomes of present-day humans from across the globe (47, 48), which made it possible to precisely estimate archaic human allele frequencies and their spatial distribution along the genome for the purposes of studying natural selection (50–52, 65). Among the many discoveries made from such “maps” of archaic human ancestry are genomic regions where archaic human DNA has been beneficial for modern humans, perhaps because it helped them adapt to new environments in Eurasia (62, 64, 66–69). Another intriguing observation was that the majority of Neanderthal DNA has been under some form of negative selection in modern human genomes (50, 51). First, two independent studies detected a striking depletion of Neanderthal DNA in a handful of extremely long segments in the genome (50, 51). The same depleted regions were later found even in populations with an additional pulse of Denisovan introgression (52, 56, 65) and became known as “deserts” of archaic ancestry (51). Because the large size of these regions is inconsistent with being a product of a simple neutral process, it has been hypothesized that they are the result of a particularly strong negative selection against some introgressed archaic human variants, possibly even drivers of “hybrid incompatibility” (50, 51). Second, it has

been shown that there is a strong negative correlation between the probability of finding an archaic human allele at any locus in the genome and the level of functional importance of that locus, with genic regions showing a particularly low proportion of surviving archaic human ancestry (51, 65). Overall, it is now clear that the fate of Neanderthal and Denisovan ancestry in modern humans has been affected by a complex interaction between various kinds of positive and negative selective forces.

Thesis outline

The aim of this thesis was to investigate the interplay between demographic factors and natural selection in ancient human introgression. In *Chapter 1*, I examine the dynamics of negative selection against Neanderthal ancestry in modern humans and establish its limits over long evolutionary timescales, particularly in the context of the decline in Neanderthal ancestry previously reported in published early modern human ancient DNA time-series data (70). In *Chapter 2*, I describe the first comprehensive analysis of the Y chromosomes of Neanderthals and Denisovans, their phylogenetic relationships to each other and to the Y chromosomes of present-day humans. I show that the Neanderthal Y chromosomes represent an extinct lineage introgressed from a population related to early modern humans, and that it is likely that these introgressed Y chromosomes completely replaced the original Y chromosomes of early Neanderthals due to their lower evolutionary fitness. Finally, in *Chapter 3* I describe a software package *admixr* for the programming language R, which facilitates the writing of automated analyses using the ADMIXTOOLS suite of population genetic programs (71). This package enabled efficient computation of large-scale admixture tests required for the work described in *Chapter 1*, but is now used in the field of ancient DNA and population genetics in general.

Chapter 1

Natural selection is a process which often occurs over long evolutionary timescales. Therefore, having temporal allele frequency data can significantly increase the statistical power to understand the dynamics of selection and estimate its parameters (72). In 2016, genome-wide ancient DNA from 51 early modern humans spanning the last 45,000 years of European prehistory was published (70), representing an important shift from studying individual ancient human genomes (60, 73, 74) towards population-scale ancient genomics. Intriguingly, the proportion of Neanderthal ancestry in individuals in this time-series showed a significant, continuous decline since the introgression, which was interpreted as direct evidence for ongoing negative selection against Neanderthal DNA in modern humans (51, 70). At the same time, two studies proposed a mechanism for the depletion of Neanderthal

ancestry observed near conserved regions in present-day human genomes in an earlier study (51, 75, 76). Their model posits that an excess of mildly deleterious Neanderthal mutations accumulated in Neanderthals as a consequence of their lower effective population size and decreased efficacy of purifying selection compared to modern humans (75, 76). Although behaving neutrally in Neanderthals, these deleterious mutations could have driven the selection against Neanderthal DNA on the modern human genetic background following the admixture with modern humans (75, 76). Coupled with this hypothesis, the empirical Neanderthal ancestry decline observed in ancient DNA time-series data (70) was so compelling that it formed the general basis of thinking about the fate of Neanderthal ancestry in modern humans (77).

In this chapter, I test the rate of Neanderthal ancestry decline observed in the early modern human ancient DNA time-series data against the predictions of population genetic theory. I show that the previously reported decline in Neanderthal ancestry is an artifact of incorrect assumptions about modern human demographic history. Using coalescent simulations, I demonstrate that gene flow from West Eurasians into Africans leads to a false inference of Neanderthal ancestry decline using previously used statistics even in the complete absence of natural selection. I propose an alternative way to estimate the proportions of Neanderthal ancestry in modern human genomes which is robust to perturbations of the underlying demographic model. Re-evaluating the levels of Neanderthal ancestry in the early modern human time-series using this robust statistic no longer shows any evidence for a decline in Neanderthal ancestry over time.

I then focus on the consequences of this result in the context of the proposed model of negative selection against accumulated deleterious Neanderthal variation in modern human genomes. I performed large-scale, forward-in-time population genetic simulations of introgression from Neanderthals into modern humans, approximating the realistic structure of the entire genome to model dynamics of selection over time as well as to track the spatial distribution of introgression along the genome. I show that the long-term trajectory of Neanderthal ancestry in modern humans remains essentially constant even under negative selection against introgression, and demonstrate that this is true regardless of the choice of model parameters or the specifics of the selection scenario. Therefore, the newly re-estimated empirical trajectory of Neanderthal ancestry in modern humans matches the predictions of population genetics theory.

It is important to note that this result does not contradict the strong signal of negative selection against introgression observed in present-day human genomes (51). On the contrary, in the selection simulations of the spatial distribution of Neanderthal ancestry along the genomes of modern humans I recover the same pattern of depletion of introgression close to functionally important genomic regions found in present-day human genomic data. This

implies that for most of the time after introgression selection acts only locally, and that the genome-wide population level of Neanderthal ancestry over time is not affected in a detectable way.

Finally, I leverage the robust estimate of Neanderthal ancestry proportions to examine the spatial heterogeneity of Neanderthal introgression in present-day human genomes. I find a much stronger depletion of Neanderthal introgression in regions affecting regulation of gene expression than in protein-coding sequences. In fact, except for a subset of strongly conserved protein-coding regions, coding sequences in general do not show any depletion in Neanderthal introgression compared to the genome-wide average. The strong signal of depletion of Neanderthal DNA in non-coding, regulatory regions is in agreement with several recent studies that show a significant influence of introgressed Neanderthal haplotypes on the regulation of gene expression in modern humans (67, 68, 78–82).

Chapter 2

Although the very first insights into deep human evolutionary history resulted from analyzing mitochondrial DNA (mtDNA) (14, 29), the most transformative recent discoveries from ancient DNA were based on autosomal sequences (34, 38, 42, 43, 70). This is largely due to the fact that unlike mtDNA, which does not recombine and provides only a limited view of history through the maternal lineage, any individual's autosomal DNA represents a rich tapestry of thousands of independent genealogies. Nevertheless, mtDNA has recently revealed intriguing discrepancies in phylogenetic relationships between archaic and modern humans. Although Neanderthals and Denisovans form a sister clade to modern humans, having split from modern humans around 500 kya, Neanderthal mtDNA sequences are much more similar to modern human mtDNA than to Denisovan mtDNA (41). Interestingly, however, some of the earliest Neanderthals from the site of Sima de los Huesos in present-day Spain carried a Denisovan-like mtDNA which is not found in later Neanderthals (83, 84). To explain these phylogenetic discrepancies, it has been suggested that the early Neanderthals carried mtDNA similar to that of Denisovans which was later replaced by gene flow from early modern humans, inferred to have occurred between 468–219 kya (84, 85). Thus, mtDNA suggests that the gene flow between Neanderthals and modern humans occurred in both directions and at much earlier time than the later Neanderthal admixture with non-Africans (84, 85).

A puzzling aspect of the Neanderthal mtDNA replacement hypothesis is the low expected probability of such mtDNA turnover given the relatively low expected rate of gene flow from modern humans into Neanderthals inferred from autosomal genomes (56, 86, 87). Furthermore, there is still a large uncertainty surrounding the timing and dynamics of the replacement process (85, 88). In this context, the Y chromosomes of Neanderthals and Denisovans could provide important insights into the history of ancient gene flow events

between different hominin groups. However, since all of the archaic human specimens which yielded sufficient amounts of sequence data were female, other than a small amount of coding sequence from a single Neanderthal (89), Y chromosomes of archaic humans have remained largely enigmatic.

In this chapter, I present the first comprehensive analysis of Y chromosomes of three Neanderthals and two Denisovans, and describe their phylogenetic relationships to each other and to the Y chromosomes of present-day humans. I show that the two Denisovan Y chromosomes fall outside of the variation of modern human Y chromosomes, and estimate that the most recent common ancestor of both groups lived around 700 kya, which is consistent with the inferred split time between Denisovan and modern human populations (39). Strikingly, all three Neanderthal Y chromosomes are more closely related to modern human Y chromosomes rather than to the Y chromosomes of Denisovans, having split from each other around 370 kya. Therefore, the phylogenetic relationships between archaic and modern human Y chromosomes recapitulate those of their mitochondrial genomes, suggesting that both mtDNA and Y chromosome gene pools of Neanderthals have been replaced by gene flow, potentially from the same population of early modern humans.

Given the low rate of early modern human gene flow estimated from nuclear genomes of Neanderthals (56, 86, 87), it is important to consider that the joint probability of the replacement of Neanderthal mtDNA and Y chromosomes by their modern human counterparts under neutrality is extremely low. However, as I show in *Chapter 1*, the assumption that introgressed DNA has behaved neutrally in an admixed modern human population is no longer tenable. Therefore, I have extended the simulations I developed for my work on the dynamics of selection against DNA introgressed from Neanderthals into modern humans (59) to study the expected behavior of a non-recombining locus introgressed in the opposite direction, i.e. from modern humans to Neanderthals. Using these simulations, I show that any factor that reduces the evolutionary fitness of Neanderthal Y chromosomes compared to modern human Y chromosomes – even by a small margin – dramatically increases the probability of the replacement of Neanderthal Y chromosomes after introgression from early modern humans. For instance, assuming introgression at 5%, 1% decrease of the fitness of Neanderthal Y chromosomes is sufficient to increase the probability of replacement to 25%, compared to only 5% probability under neutrality. Assuming a 2% decrease of Neanderthal Y chromosome fitness increases the replacement probability to 50%. Therefore, a simple nearly-neutral model of introgression from a population with a larger historical effective population size into a smaller population does, in fact, predict a rapid spread of the introgressed non-recombining locus through the admixed population.

Chapter 3

With the advent of large-scale genomic studies (34, 35, 47, 48), the field of human genetics has become increasingly computational. Moreover, in addition to a growing emphasis on efficient software tools, the recent debate surrounding the “reproducibility crisis” in science (90–93) has highlighted the need for these tools to facilitate reproducible research with as little human intervention and as much automation as possible (94, 95).

One of the most important pieces of software in the field of ancient DNA and human evolutionary genetics in general is ADMIXTOOLS (71). ADMIXTOOLS is a collection of command-line programs for calculating various population genetic statistics and for testing admixture hypotheses, implementing f_3 , f_4 , D , and f_4 -ratio statistics, as well as methods such as *qpWave* and *qpAdm*. Since its publication in 2012, this software has been used in nearly all recent ancient human DNA studies (70, 96–103), and has become a standard tool for population genetics analyses in other species as well (104–106). Although powerful and comprehensive, ADMIXTOOLS programs require a significant amount of manual work to perform even the most basic analyses. Historically, this has not been a critical issue as the number of ancient human samples in population genetic studies was quite limited (60, 74, 107). However, as the number of ancient genomes continues to rise rapidly (99, 103, 108), performing computational analyses which are fully reproducible becomes challenging (94, 95, 109).

As the work I described in *Chapter 1* (59) involved performing thousands of admixture tests on hundreds of samples using the ADMIXTOOLS programs, I developed a programmable, easy-to-use interface for testing population genetic hypotheses, and for estimating admixture parameters using the ADMIXTOOLS suite as a backend. The software, called *admixr*, is implemented as a package for the programming language R, which is a *de facto* standard framework for statistical analyses and plotting (110). This package allows performing all steps of ADMIXTOOLS analyses directly in R, eliminating any need for manual work and error-prone scripting, yet it still internally leverages the efficient and well-tested ADMIXTOOLS algorithms. Most importantly, in order to help the field of ancient DNA move towards more reproducible analyses, *admixr* is explicitly designed with a complete automation and full reproducibility in mind (94, 95, 109). In the year since its publication (109), the package has been adopted for population genetic research in humans and other species (111–122).

Conclusions

The work described in this thesis has contributed to several aspects of the study of ancient human introgression, both from a demographic perspective and also from the viewpoint of natural selection.

First, I have shown that the proportion of Neanderthal ancestry in modern humans has remained essentially constant since the introgression, and that this is consistent with the expectation dictated by population genetic theory. This corrected the widely accepted notion that the levels of Neanderthal ancestry in Eurasians declined monotonically from around 6% at the time of introgression to approximately 1-2% in people today (70). I have demonstrated that the most likely cause of the false inference of Neanderthal decline is an unaccounted for gene flow between West Eurasians and Africans. The newly established, stable trajectory of Neanderthal ancestry can now serve as a constraint for future models of selection acting on archaic human introgression. It also suggests that the previously hypothesized dilution of Neanderthal ancestry in Europeans via gene flow from a population with little to no Neanderthal introgression (53), if it occurred at all, must have had only a limited effect. In general, my findings highlight the need for careful examination of assumptions behind population admixture estimates beyond simple intuitive reasoning, ideally through exhaustive simulations which are now trivial to perform even for genome-scale sized data sets (123). Furthermore, my results point to a promising avenue of future research into the exact timing, geographical extent, and source of the gene flow from West Eurasia into Africa, leveraging the dense spatial and temporal ancient DNA sampling of various West Eurasian populations. For instance, a recent study has focused on the impact of this gene flow in greater detail and demonstrated that it can account for a majority of the apparent differences in Neanderthal ancestry between Europe and East Asia (56). Finally, I have shown that there is an unbiased way to estimate Neanderthal ancestry proportions, which will be particularly important as DNA from individuals who lived close to the time of Neanderthal introgression becomes available (124). Another outcome of this work is the R package *admixr*, which allows programming of large-scale, reproducible population genetic analyses of thousands of present-day and ancient samples.

Second, I have presented the first comprehensive analysis of Denisovan and Neanderthal Y chromosomes. I have shown that their phylogenetic relationships to modern human Y chromosomes mirror that of their mitochondrial genomes, with Neanderthal Y chromosomes being more closely related to modern humans than to Denisovans. This suggests that similarly their mitochondrial DNA, Y chromosomes of Neanderthals were completely replaced via gene flow from a population related to early modern humans. This work adds to the growing body of evidence for ancient gene flow from modern humans into

Neanderthals (85–87) and, in particular, to the evidence for modern human presence outside of Africa at least a hundred thousand years prior to the major out of Africa migration (125, 126). Because all three Neanderthal Y chromosomes which I have analysed come from West Eurasia, a more extensive geographical sampling will be needed to resolve the extent of the Y chromosome replacement. However, given that the signal of mtDNA replacement extends over the entire Eurasian continent (85) and given the evidence for continental-scale migrations of Neanderthals in general (88, 127), it is reasonable to assume that the replacement of Neanderthal Y chromosomes reached even the easternmost corner of their habitat in Siberia. Similarly, sampling the Y chromosomes of older Neanderthals will help to establish the timing of the replacement dynamics and narrow down the selection parameters driving the replacement process. In particular, based on the age of the Sima de los Huesos individuals and the phylogenetic status of their mtDNA (83, 84), their Y chromosomes should cluster with the Denisovan Y chromosomes. If more sequence data from the Sima de los Huesos hominins shows this to be the case, their Y chromosomes could provide a boundary for establishing when the gene flow into Neanderthals occurred. Finally, although my simulations of the replacement dynamics of Neanderthal Y chromosomes show that a simple model of weaker purifying selection on the Neanderthal lineage is sufficient to explain the observed turnover of Y chromosomes on its own, the exact functional mechanism which drove the replacement remains unclear. The factors which negatively affect male fertility are numerous and complex and many of them involve structural changes and chromosomal rearrangements (128), none of which have been possible to identify because of the limits of quality of the available ancient DNA. However, future technological advancements in ancient DNA extraction and sequencing technology will hopefully yield multiple high-coverage Y chromosomes of Neanderthals and Denisovans without resorting to DNA capture. In that case, *de novo* assembly of Neanderthal and Denisovan Y chromosomes and their comparisons to modern human Y chromosomes could determine which changes may have led to the lower fitness of Neanderthal Y chromosomes and, consequently, to their eventual replacement.

Zusammenfassung

Einleitung

1856 wurde im Neandertal im heutigen Deutschland ein ungewöhnliches Skelett entdeckt. Dieses Individuum, das später als das Typusexemplar *Feldhofer 1* bezeichnet wurde, gehörte, wie viele andere, die in den Jahrzehnten danach gefunden wurden, zu einer Hominin-Gruppe namens Neandertaler (1). Aus morphologischer Sicht trugen diese Menschen zweifellos ähnliche Merkmale wie wir, und doch wiesen sie Merkmale sowie kulturelle Praktiken auf, die sich vom modernen Menschen deutlich genug unterschieden, um in den folgenden mehr als hundert Jahren heftige Debatten zu entfachen. Waren die Neandertaler nur ein Übergangszustand auf dem Weg zum modernen menschlichen Phänotyp oder stellten sie eine völlig neue Spezies von Menschen dar (2)?

Im Laufe der Jahrzehnte wurde immer deutlicher, dass beide Gruppen unabhängig voneinander entwickelte, zeitgleiche Abstammungslinien repräsentieren (3, 4). Tatsächlich verschwanden die Neandertaler, die weite Teile des eurasischen Kontinents (5) besiedelt hatten, nachdem sie vor etwa 400.000 Jahren in Europa erstmals aufgetaucht waren (6, 7), vor etwa 40.000 Jahren aus den Fossilienfunden (8), mehrere tausend Jahre nach der Ankunft der ersten modernen Menschen in Eurasien (9–13). Damit ist zu einer der wichtigsten Fragen der Anthropologie geworden, inwieweit die Neandertaler, wenn überhaupt, mit dem modernen Menschen während ihrer jahrtausende langen Überlappung in Eurasien interagiert haben. Sind die Neandertaler vollständig durch den modernen Menschen ersetzt worden, ohne eine Spur zu hinterlassen, als der moderne Mensch aus Afrika auswanderte, um die Welt zu kolonisieren (14)? Haben sich die beiden Menschengruppen vermischt und Nachkommen gezeugt? Könnten wir Spuren des genetischen Erbguts der Neandertaler in den modernen menschlichen Genomen nachweisen?

Morphologische Analysen einiger der frühesten modernen menschlichen Fossilien, die in Europa gefunden wurden, ergaben Hinweise auf morphologische Merkmale, die von Neandertalern stammen, was darauf hindeutet, dass einige Vermischungen stattgefunden haben (9, 15–21), obwohl der "Hybrid"-Status einiger dieser Individuen zwischen Neandertalern und modernen Menschen in Frage gestellt wurde (22–24). Darüber hinaus deuteten andere Hinweise auf die Möglichkeit nicht nur eines genetischen, sondern auch eines kulturellen Austauschs zwischen Neandertalern und modernen Menschen hin (25). Letztlich bedurfte es jedoch eines genetischen Arguments, um die Frage zu klären, ob die Neandertaler DNA zum modernen Menschen beigetragen haben und ob dieses genetische Erbe bei den heute lebenden Menschen immer noch vorhanden ist.

Die erste Studie über antike DNA wurde 1984 veröffentlicht und zeigte, dass DNA nicht nur in einem Museumsexemplar eines ausgestorbenen Quagga fast 150 Jahre nach dessen Tod überlebt hat, sondern dass sie durch molekulares Klonen erfolgreich amplifiziert und

sequenziert werden konnte (26). Ein Jahr später zeigte die DNA-Extraktion aus ägyptischen Mumien, dass dasselbe auch für mehrere tausend Jahre ältere Museumsexemplare gelten könnte (27). Die wahre Revolution auf dem Gebiet der alten DNA kam jedoch mit der Entwicklung der Polymerase-Kettenreaktion (PCR) (28), mit der die erste DNA-Sequenz eines Neandertalers bestimmt werden konnte (29).

Diese Sequenz, die zum *Feldhofer 1* Typusexemplar gehörte, repräsentierte 379 Basenpaare der mitochondrialen DNA (mtDNA)-Kontrollregion, und ihr phylogenetischer Vergleich mit modernen menschlichen mtDNAs zeigte, dass die Neandertaler-mtDNA vor der Diversifizierung der mtDNA-Linien heute lebender Menschen divergierte (29). Die Sequenzierung kompletter mtDNA-Genome vieler weiterer Neandertaler im Laufe des folgenden Jahrzehnts kam immer zum gleichen Ergebnis (5, 30–32) und unterstützte im Rahmen der breiteren anthropologischen Debatte die Hypothese der Ersetzung der Neandertaler durch den modernen Menschen ohne jegliche Vermischung (29). Eine theoretische Analyse der ersten mtDNA-Studie bei Neandertalern (29) warnte jedoch davor, dass, obwohl die phylogenetischen Beziehungen der mtDNA Modelle einer zufälligen Paarung zwischen Neandertalern und modernen Menschen ablehnen, es jedoch andere Szenarien des Genflusses gibt, die in Betracht gezogen werden müssen (33). Unter diesen konnten nur die Modelle abgelehnt werden, in denen sehr großen Mengen an mtDNA des Neandertalers an den modernen Menschen weitergegeben wird, da die verfügbaren Daten nur eine begrenzte statistische Aussagekraft hatten und die mtDNA allein einen engen Blick auf die menschliche Evolutionsgeschichte ermöglicht (30, 33). Um das Problem des Neandertaler-Genflusses beim modernen Menschen vollständig zu lösen, waren Informationen aus dem nuklearen Genom erforderlich, das eine Zusammensetzung aus Tausenden von Genealogien der Vorfahren eines jeden Individuums darstellt.

Die erste Analyse des Entwurfs des nuklearen Genoms eines Neandertalers wurde 2010 veröffentlicht (34). Sie wurde weniger als zehn Jahre nach der Zusammenstellung des ersten vollständigen Entwurfs des Genoms des modernen Menschen (35) abgeschlossen und stellte eine bemerkenswerte technische Errungenschaft dar, die einen großen koordinierten Aufwand erforderte (36, 37). Der Vergleich des Neandertaler-Genoms mit den Genomen heutiger Menschen zeigte, dass Neandertaler mit heutigen Nicht-Afrikanern mehr Allele teilen als mit Afrikanern, und dass der Großteil dieses Signals vom Genfluss der Neandertaler in die Vorfahren der Nicht-Afrikaner stammen muss (34). Spätere technische Entwicklungen ermöglichten die Sequenzierung der Genome von drei weiteren Neandertalern mit hoher Qualität, was dazu führte, dass der Anteil der Neandertaler-Abstammung bei heutigen Nicht-Afrikanern auf etwa 2% geschätzt wurde und die abgeleitete Zeit der Aufspaltung zwischen Neandertalern und den Vorfahren des modernen Menschen zwischen 520.000 und 630.000 Jahren lag (38–40).

Der definitive genetische Nachweis der Introgression von Neandertalern in den modernen Menschen löste eines der am meisten diskutierten Probleme in der evolutionären Anthropologie (34). Die alte DNA hat jedoch weiterhin neue Erkenntnisse über die menschliche Evolutionsgeschichte geliefert. Im Jahr 2010 enthüllte die aus einer distalen Handphalanx extrahierte DNA, die in der Denisova-Höhle in Sibirien ausgegraben wurde, einen unbekanntem Typ des mitochondrialen Genoms, der von einem gemeinsamen Vorfahren der mtDNA-Sequenzen von Neandertalern und modernen Menschen vor etwa 1 Million Jahren divergiert war, was die Möglichkeit aufkommen ließ, dass sein Träger zu einer völlig neuen Hominin-Gruppe gehörte (41). Tatsächlich wurde sie bei der Sequenzierung des nuklearen Genoms dieses Individuums einer noch unbekanntem Hominin-Population zugeordnet, die eine Schwestergruppe der Neandertaler war (38, 42, 43). Bemerkenswerterweise wurde festgestellt, dass diese Population, die Denisova-Menschen genannt wurde, bis zu 6% zur Abstammung der Menschen im heutigen Papua-Neuguinea und der Inselbewohner von Bougainville beigetragen hat (42, 43). Eine spätere Analyse der geografischen Verteilung der Abstammung der Denisova-Menschen in einer größeren Stichprobe südostasiatischer und ozeanischer Populationen zeigte, dass der Genfluss von Denisova-Menschen wahrscheinlich in Südostasien stattgefunden hat (44). Da Denisova-Menschen bisher nur in Sibirien und auf dem tibetischen Plateau gefunden wurden (45), bedeutet dies, dass sie eine außergewöhnlich große geographische Region von Sibirien bis zum tropischen Asien besetzt haben müssen (44).

Nach dem Nachweis, dass sich moderne Menschen mit archaischen Menschen bei ihrer Migration aus Afrika vermischt haben (34, 42, 46), verlagerte sich der Schwerpunkt der Untersuchung der Introgression archaischer Menschen auf differenziertere Fragen. Wann und wo fand die Vermischung statt? Wie häufig kam es zu Interaktionen zwischen archaischen und modernen Menschen? Gab es funktionelle Konsequenzen der eingeführten DNA archaischer Menschen auf die Physiologie des modernen Menschen?

Aufgrund gleicher Ähnlichkeit des Neandertaler-Genoms zu den Genomen mehrerer Europäer und Ostasiaten kam der erste Entwurf einer Neandertaler-Genomstudie zu dem Schluss, dass der Genfluss wahrscheinlich nach der Migration aus Afrika, aber vor der Aufspaltung der heutigen nicht-afrikanischen Populationen im Nahen Osten stattgefunden hat (34). Eine spätere Analyse der Längenverteilung der Neandertaler-Haplotypen bestätigte den Zeitpunkt der Introgression vor etwa 55.000 Jahren (46). Neuere Analysen der geografischen Verteilung des Anteils der Neandertaler-Abstammung in Genomen heute lebender Menschen aus aller Welt (43, 47, 48) ergaben allerdings ein komplexeres Bild der Geschichte der Neandertaler-DNA in unseren Genomen. Vor allem scheinen ostasiatische Populationen etwa 20% mehr Neandertaler-DNA zu tragen als Europäer (39, 43, 49–52).

Dieses Übermaß an Neandertaler-Abstammung bei Menschen Ostasiens wird im Wesentlichen durch die folgenden drei sich im Wettstreit miteinander befindlichen Hypothesen zu erklären versucht. Zum einen könnten sich die Vorfahren heutiger Ostasiaten nach der Trennung von den Vorfahren heutiger Europäer ein zweites Mal mit Neandertalern vermischt haben, was zu einem zusätzlichen Anteil an Neandertaler-DNA führen würde (43). Alternativ dazu könnte nach der Introgression in die Vorfahren aller Nicht-Afrikaner der Neandertaler-Anteil in den Genomen heutiger Europäer durch Vermischung mit einer anderen, möglicherweise afrikanischen, Population reduziert worden sein, die wenig bis gar keine Neandertaler-DNA trug (43, 53). Und schließlich wurde die bei heutigen Ostasiaten beobachtete größere Menge an Neandertaler-DNA als Folge einer schwächeren Selektion gegen Neandertaler-DNA im Vergleich zu Europäern angeführt, was auf eine geringere effektive Populationsgröße bei den Vorfahren heutiger Ostasiaten zurückgeführt werden kann (51). Durch populationsgenetische Modellierung des Zusammenspiels von Selektionseffizienz und genetischen Drifts kann das Selektionsmodell nahezu ausgeschlossen werden. Ein zweiter "Puls" von Neandertaler-Introgression in die Vorfahren heutiger Ostasiaten oder das "Verdünnungs-szenario" erscheinen in diesen Modellen plausibler (50, 54, 55).

Des weiteren hat eine kürzlich durchgeführte Studie gezeigt, dass das Übermaß an Neandertaler-DNA bei heutigen Ostasiaten im Vergleich zu Europäern zu weiten Teilen auf ein statistisches Artefakt zurückzuführen sein könnte (56). Gängige Methoden zum Nachweis von introgressierter Neandertaler-DNA basieren auf der Annahme, dass bestimmte afrikanische Gruppen überhaupt keine Neandertaler-DNA tragen, und nutzen diese als "Referenzlinie" für den Nachweis von Neandertalersegmenten in der DNA anderer Populationen. (50, 51, 56, 57). Die meisten afrikanischen Populationen scheinen jedoch einen Teil Neandertaler-DNA in sich zu tragen, die nicht aus Neandertaler-Introgression, sondern aus Rückmigration von Menschen aus Eurasien nach Afrika stammt (58, 59). Da dieser Rückfluss nach aktuellem Kenntnisstand wahrscheinlich aus einer westeurasischen Quelle stammt, möglicherweise sogar von Vorfahren heutiger Europäer selbst, kann dies zu einem Verzerrungseffekt gegen die Entdeckung von Neandertaler-DNA in den Genomen heutiger Europäer führen (56, 59). Bei der Anwendung von Methoden, die keine expliziten Annahmen über die Komplexität der demographischen Geschichte moderner Menschen treffen, entfällt das Signal des vermeintliche Übermaßes von Neandertaler-DNA bei Ostasiaten weitgehend, wodurch die Notwendigkeit für eine "Verdünnungshypothese" oder der Hypothese der mehrfachen Introgressionsereignisse verringert wird (56). Meine eigene Arbeit, die in *Kapitel 1* ausführlich beschrieben wird, hat neue Belege für den Genfluss von Westurasien nach Afrika geliefert und die Notwendigkeit robuster Methoden zur Bestimmung von Neandertaler-DNA in den Genomen moderner Menschen unterstrichen (59). Die Frage, wie viele Introgressionspulse vom Neandertaler (oder Denisova-Menschen) genau während der Zeit der Überschneidung von archaischen und modernen Menschen in Eurasien aufgetreten sind,

bleibt vorerst unbeantwortet. Hinweise auf multiple Introgressionsereignisse werden jedoch immer deutlicher (12, 52, 60, 61).

Neben der Aufschlüsselung einiger Aspekte vergangener Begegnungen von archaischen und modernen Menschen aus demographischer Sicht, hat die alte DNA zahlreiche Einblicke in funktionelle Konsequenzen der Introgression der DNA von Neandertaler und Denisova-Menschen auf die moderne menschliche Biologie geliefert (62–64). Dies ist zum Teil den umfangreichen Datenbanken mit Tausenden von Genomen heutiger Menschen aus aller Welt zu verdanken (47, 48), die es ermöglichen, Frequenzen "archaischer Allele" und ihre räumliche Verteilung im Genom präzise abzuschätzen, um die dort wirkende natürliche Selektion zu untersuchen (50–52, 65). Unter den Erkenntnissen, die sich aus der Erstellung solcher "Landkarten" archaisch-menschlicher DNA in den Genomen moderner Menschen ergeben, ist die Entdeckung von Genomregionen, in denen archaische-menschliche DNA für den modernen Menschen von Vorteil war - vielleicht, weil sie ihm bei der Anpassung an die eurasische Umwelt geholfen hat (62, 64, 66–69). Eine weitere faszinierende Beobachtung war, dass die Mehrheit der Neandertaler-DNA in den modernen menschlichen Genomen in irgendeiner Form einer negativen Selektion unterzogen worden ist (50, 51). Zunächst wurde in zwei unabhängigen Studien eine auffällige Abnahme der Neandertaler-DNA in einer Handvoll extrem langer Abschnitte des Genoms festgestellt (50, 51). Dieselben verarmten Regionen wurden später sogar in Populationen mit einem zusätzlichen Puls der Introgression von Denisova-Menschen gefunden (52, 56, 65) und wurden als "Wüsten" archaischer Abstammung bekannt (51). Da die Größe dieser Regionen nicht damit vereinbar ist, ein Produkt eines einfachen neutralen Prozesses zu sein, wurde die Hypothese aufgestellt, dass sie das Ergebnis einer besonders starken Negativselektion gegenüber einigen introgressierten archaischen menschlichen Varianten sind, möglicherweise sogar Treiber einer "Hybridinkompatibilität" (50, 51). Zweitens hat sich gezeigt, dass es eine starke negative Korrelation zwischen der Wahrscheinlichkeit, ein archaisches menschliches Allel an einem beliebigen Locus im Genom zu finden, und dem Grad der funktionellen Bedeutung dieses Locus gibt, wobei die Genregionen einen besonders geringen Anteil an überlebenden archaischen menschlichen Abstammungslinien aufweisen (51, 65). Insgesamt ist nun klar, dass das Schicksal der Abstammung vom Neandertaler und von Denisova-Menschen im modernen Menschen durch ein komplexes Zusammenspiel verschiedener Arten positiver und negativer selektiver Kräfte beeinflusst wurde.

Gliederung der Thesis

Ziel dieser Arbeit war die Untersuchung des Zusammenspiels zwischen demographischen Faktoren und natürlicher Selektion in alten Introgressionsereignissen im Menschen. In *Kapitel 1* untersuche ich die Dynamik der negativen Selektion gegen die Neandertaler-Abstammung beim modernen Menschen und stelle ihre Grenzen über lange evolutionäre Zeiträume fest, insbesondere im Zusammenhang mit dem Rückgang der Neandertaler-Abstammung, über den bereits in publizierten Zeitreihendaten von alter DNA von frühen modernen Menschen berichtet wurde (70). In *Kapitel 2* beschreibe ich die erste umfassende Analyse von Y-Chromosomen von Neandertalern und Denisova-Menschen, ihre phylogenetischen Beziehungen zueinander und zu den Y-Chromosomen des heutigen Menschen. Ich zeige, dass die Y-Chromosomen der Neandertaler eine ausgestorbene Abstammungslinie repräsentieren, die aus einer mit dem frühen modernen Menschen verwandten Population eingeführt geworden ist, und dass es wahrscheinlich ist, dass diese introgressierten Y-Chromosomen die ursprünglichen Y-Chromosomen der frühen Neandertaler aufgrund ihrer geringeren evolutionären Fitness vollständig ersetzt haben. Abschließend beschreibe ich in *Kapitel 3* ein Softwarepaket *admixr* für die Programmiersprache R, welches das Schreiben automatisierter Analysen mit der ADMIXTOOLS-Suite populationsgenetischer Programme erleichtert (71). Dieses Paket ermöglichte die effiziente Berechnung von groß angelegten Vermischungsanalysen, die für die in *Kapitel 1* beschriebenen Arbeiten erforderlich sind, wird aber heute im Bereich der alten DNA und der Populationsgenetik im Allgemeinen weit verbreitet eingesetzt.

Kapitel 1

Die natürliche Selektion ist ein Prozess, der sich oft über lange evolutionäre Zeiträume erstreckt. Daher kann die Verfügbarkeit zeitlicher Allelhäufigkeitsdaten die statistische Aussagekraft zum Verständnis der Dynamik der Selektion und zur Schätzung ihrer Parameter signifikant erhöhen (72). Im Jahr 2016 wurde die genomweite alte DNA von 51 frühen modernen Menschen aus den letzten 45.000 Jahren der europäischen Vorgeschichte veröffentlicht (70), was eine wichtige Verschiebung von der Untersuchung individueller alter menschlicher Genome (60, 73, 74) hin zu einer bevölkerungsweiten alten Genomik darstellt. Interessanterweise zeigte der Anteil der Abstammung vom Neandertaler bei den Individuen in dieser Zeitreihe einen signifikanten, kontinuierlichen Rückgang seit der Introgression, was als direkter Beweis für eine anhaltende negative Selektion gegen Neandertaler-DNA beim modernen Menschen interpretiert wurde (51, 70). Unabhängig voneinander schlugen zwei Studien einen Mechanismus für die Reduzierung der Neandertaler-Abstammung vor, die in

einer früheren Studie in der Nähe von konservierten Regionen im Genom von heute lebenden Menschen beobachtet wurde (51, 75, 76). Ihr Modell postuliert, dass sich leicht schädliche Neandertaler-Mutationen bei Neandertalern als Folge ihrer geringeren effektiven Populationsgröße und der verminderten Wirksamkeit der purifizierender Selektion akkumulierten (75, 76). Obwohl sie sich bei Neandertalern nach der Vermischung mit modernen Menschen neutral verhalten, könnten diese schädlichen Mutationen die Selektion gegen Neandertaler-DNA bei einem modernen humangenetischen Hintergrund vorangetrieben haben (75, 76). Diese Hypothese war in Verbindung mit dem empirischen Rückgang der Neandertaler-Abstammung, der in Zeitreihendaten alter DNA beobachtet wurde (70), so überzeugend, dass sie die allgemeine Grundlage bildete, um über das Schicksal der Neandertaler-Abstammung beim modernen Menschen nachzudenken (77).

In diesem Kapitel teste ich die Rate des Rückgangs der Neandertaler-Abstammung, die in den Zeitreihendaten alter DNA früherer moderner Menschen beobachtet wurde, gegen die Vorhersagen der Theorien der Populationsgenetik. Ich zeige, dass der zuvor berichtete Rückgang der Neandertaler-Abstammung ein Artefakt falscher Annahmen über die Populationsgeschichte moderner Menschen ist. Mit Hilfe von Koaleszenz-Simulationen zeige ich, dass der Genfluss von Westeurasiern zu Afrikanern zu einer falschen Schlussfolgerung über den Rückgang der Neandertaler-Abstammung selbst bei völliger Abwesenheit natürlicher Selektion führt, wenn zuvor verwendete Statistiken herangezogen werden. Ich schlage eine alternative Methode zur Schätzung des Anteils der Neandertaler-Abstammung im Genom von modernen Menschen vor, die robust gegenüber Störungen des zugrunde liegenden demographischen Modells ist. Eine Neubewertung der Höhe der Neandertaler-Abstammung in den Zeitreihen früherer moderner Menschen unter Verwendung dieser robusten Statistik zeigt keine Anzeichen mehr für einen Rückgang der Neandertaler-Abstammung über den Lauf der Zeit.

Anschließend konzentriere ich mich auf die Folgen dieses Ergebnis im Zusammenhang mit dem vorgeschlagenen Modell der negativen Selektion gegen die akkumulierte schädliche Neandertaler-Variationen im Genom moderner Menschen. Ich führte groß angelegte vorwärtsgerichtete populationsgenetische Simulationen der Introgression von Neandertalern in moderne Menschen durch, wobei ich die realistische Struktur des gesamten Genoms näherte, um die Dynamik der Selektion über die Zeit zu modellieren und die räumliche Verteilung der Introgression entlang des Genoms zu verfolgen. Ich zeige, dass die langfristige Trajektorie der Neandertaler-Abstammung beim modernen Menschen auch bei negativer Selektion gegen Introgression im Wesentlichen konstant bleibt, und zeige, dass dies unabhängig von der Wahl der Modellparameter oder den Besonderheiten des Selektionsszenarios zutrifft. Daher stimmt die neu abgeschätzte empirische Trajektorie der

Neandertaler-Abstammung beim modernen Menschen mit den Vorhersagen der Theorie der Populationsgenetik überein.

Es ist wichtig anzumerken, dass dieses Ergebnis nicht im Widerspruch zu dem starken Signal der negativen Selektion gegen Introgression steht, das in Genomen heute lebender Menschen beobachtet wird (51). Im Gegenteil, in den Selektionssimulationen der räumlichen Verteilung der Neandertaler-Abstammung entlang der Genome des modernen Menschen stelle ich dasselbe Muster der Verringerung von Introgression in der Nähe funktionell wichtiger genomischer Regionen fest, welches in Genomdaten heutiger Menschen zu finden ist. Dies impliziert, dass die Selektion nach der Introgression die meiste Zeit über nur lokal wirkt und dass auf Populationsebene die genomweite Höhe an Neandertaler-Abstammung im Laufe der Zeit nicht in nachweisbarer Weise beeinflusst wird.

Schließlich nutze ich die robuste Schätzung der Proportionen der Neandertaler-Abstammung, um die räumliche Heterogenität der Neandertaler-Introgression in den Genomen heutiger Menschen zu untersuchen. Ich finde eine viel stärkere Abnahme der Neandertaler-Introgression in Regionen, die die Regulation der Genexpression beeinflussen, als in proteinkodierenden Sequenzen. Tatsächlich zeigen kodierende Sequenzen im Allgemeinen, abgesehen von einer Untergruppe stark konservierter proteinkodierender Regionen, im Vergleich zum genomweiten Durchschnitt keine Verringerung der Neandertaler-Introgression. Das starke Signal der Verringerung von Neandertaler-DNA in nicht-kodierenden, regulatorischen Regionen steht im Einklang mit mehreren neueren Studien, die einen signifikanten Einfluss introgressiver Neandertaler-Haplotypen auf die Regulation der Genexpression beim modernen Menschen zeigen (67, 68, 78–82).

Kapitel 2

Obwohl die allerersten Einblicke in die tiefe menschliche Evolutionsgeschichte aus der Analyse der mitochondrialen DNA (mtDNA) resultierten (14, 29), basierten die neueren Entdeckungen, die das Feld der alter DNA am meisten transformierten haben, auf autosomalen Sequenzen (34, 38, 42, 43, 70). Dies ist weitgehend darauf zurückzuführen, dass die autosomale DNA eines Individuums im Gegensatz zur mtDNA, welche nicht rekombiniert und daher nur einen begrenzten Blick auf die Geschichte entlang der mütterlichen Abstammungslinie erlaubt, einen reichen Teppich aus Tausenden von unabhängigen Genealogien darstellt. Nichtsdestotrotz hat die mtDNA vor kurzem faszinierende Diskrepanzen in den phylogenetischen Beziehungen zwischen archaischen und modernen Menschen aufgedeckt. Obwohl Neandertaler und Denisova-Menschen eine Schwestergruppe des modernen Menschen bilden, die sie sich vom modernen Menschen etwa vor 500.000 Jahren abgespalten haben, sind die mtDNA-Sequenzen der Neandertaler der mtDNA moderner Menschen viel ähnlicher als der mtDNA der Denisova-Menschen (41).

Interessanterweise trugen jedoch einige der frühesten Neandertaler vom Fundort Sima de los Huesos im heutigen Spanien eine Denisova-ähnliche mtDNA, die bei späteren Neandertalern nicht gefunden wird (83, 84). Um diese phylogenetischen Diskrepanzen zu erklären, wurde angenommen, dass die frühen Neandertaler eine mtDNA trugen, die derjenigen der Denisova-Menschen ähnlich war, die später durch einen Genfluss von frühen modernen Menschen ersetzt wurde, von dem angenommen wird, dass er vor zwischen 468.000 und 219.000 Jahren stattgefunden hat (84, 85). Somit hat die mtDNA gezeigt, dass der Genfluss zwischen Neandertalern und modernen Menschen in beide Richtungen und zu einem viel früheren Zeitpunkt stattgefunden hat als die spätere Vermischung von Neandertalern mit Nicht-Afrikanern (84, 85).

Ein rätselhafter Aspekt der Hypothese über den Austausch der Neandertaler-mtDNA ist die geringe zu erwartende Wahrscheinlichkeit eines solchen mtDNA-Austauschs angesichts der relativ geringen zu erwarteten Rate des Genflusses vom modernen Menschen in den Neandertaler, die von autosomalen Genomen abgeleitet wird (56, 86, 87). Darüber hinaus besteht nach wie vor eine große Unsicherheit bezüglich des Zeitpunkts und der Dynamik des Austauschprozesses (85, 88). In diesem Zusammenhang könnten die Y-Chromosomen von Neandertalern und Denisova-Menschen wichtige Erkenntnisse über die Geschichte alter Genflussereignisse zwischen verschiedenen Hominingruppen liefern. Da jedoch alle Proben von archaischen Menschen, die ausreichende Mengen an Sequenzdaten lieferten, weiblich waren, abgesehen von einer kleinen Menge kodierender Sequenzen eines einzelnen Neandertalers (89), sind die Y-Chromosomen von archaischen Menschen weitgehend rätselhaft geblieben.

In diesem Kapitel stelle ich die erste umfassende Analyse der Y-Chromosomen von drei Neandertalern und zwei Denisova-Menschen vor und beschreibe ihre phylogenetischen Beziehungen zueinander und zu den Y-Chromosomen heutiger Menschen. Ich zeige, dass die beiden Y-Chromosomen der Denisova-Menschen außerhalb der Variation der Y-Chromosomen des modernen Menschen liegen, und schätze, dass der jüngste gemeinsame Vorfahre beider Gruppen vor etwa 700.000 Jahren lebte, was mit der abgeleiteten Zeitspanne für die Abspaltung zwischen Denisova-Menschen und moderner Menschen übereinstimmt (39). Auffallend ist, dass alle drei Y-Chromosomen der Neandertaler eher mit den Y-Chromosomen des modernen Menschen als mit den Y-Chromosomen der Denisova-Menschen verwandt sind, da sie sich vor etwa 370.000 Jahren voneinander getrennt haben. Daher rekapitulieren die phylogenetischen Beziehungen zwischen archaischen und modernen menschlichen Y-Chromosomen die ihrer mitochondrialen Genome, was darauf hindeutet, dass sowohl die mtDNA- als auch die Y-Chromosomen-Genpools der Neandertaler durch Genfluss ersetzt worden sind, möglicherweise aus derselben Population früher moderner Menschen.

Angesichts der niedrigen Rate des Genflusses von frühen modernen Menschen, die aus den nuklearen Genomen von Neandertalern geschätzt wird (56, 86, 87), ist es wichtig zu berücksichtigen, dass die gemeinsame Wahrscheinlichkeit des Austauschs der mtDNA und der Y-Chromosomen der Neandertaler durch ihre Gegenstücke von moderner Menschen unter Neutralität extrem gering ist. Wie ich jedoch in *Kapitel 1* zeige, ist die Annahme, dass sich introgressierte DNA in einer gemischten Population neutral verhalten hat, nicht mehr haltbar. Daher habe ich die Simulationen, die ich für meine Arbeit über die Dynamik der Selektion gegen die vom Neandertaler in den modernen Menschen introgressierte DNA entwickelt habe (59), erweitert, um das zu erwartende Verhalten eines nicht rekombinierenden Locus zu untersuchen, der in die entgegengesetzte Richtung introgressiert ist, d.h. vom modernen Menschen zum Neandertaler. Anhand dieser Simulationen zeige ich, dass jeder Faktor, der die evolutionäre Fitness der Y-Chromosomen von Neandertalern im Vergleich zu den Y-Chromosomen des modernen Menschen reduziert - und sei es auch nur geringfügig - die Wahrscheinlichkeit des Ersatzes der Y-Chromosomen von Neandertalern nach der Introgression vom frühen modernen Mensch dramatisch erhöht. So reicht beispielsweise die Annahme einer Introgression von 5% und 1% Abnahme der Fitness der Neandertaler-Y-Chromosomen aus, um die Wahrscheinlichkeit des Austauschs auf 25% zu erhöhen, während sie bei Neutralität nur 5% beträgt. Die Annahme einer Abnahme der Fitness der Y-Chromosomen des Neandertalers um 2% erhöht die Austauschwahrscheinlichkeit auf 50%. Daher sagt ein einfaches, nahezu neutrales Modell der Introgression von einer Population mit einer größeren historischen effektiven Populationsgröße in eine kleinere Population tatsächlich eine rasche Ausbreitung des introgressierten nicht-rekombinierenden Locus durch die gemischte Population voraus.

Kapitel 3

Mit dem Aufkommen von groß angelegten Genomstudien (34, 35, 47, 48), ist das Feld der Humangenetik zunehmend rechenintensiv geworden. Darüber hinaus hat die jüngste Debatte über die "Reproduzierbarkeitskrise" in der Wissenschaft (90–93) neben der wachsenden Bedeutung effizienter Software-Werkzeuge auch die Notwendigkeit dieser Werkzeuge hervorgehoben, um reproduzierbare Forschung mit so wenig menschlichen Eingriffen und so viel Automatisierung wie möglich zu ermöglichen (94, 95).

Eine der wichtigsten Software auf dem Gebiet der alten DNA und der menschlichen Evolutionsgenetik im Allgemeinen ist ADMIXTOOLS (71). ADMIXTOOLS ist eine Sammlung von Kommandozeilenprogrammen zur Berechnung verschiedener populationsgenetischer Statistiken und zur Prüfung von Vermischungshypothesen, indem sie die f_3 , f_4 , D , und f_4 -Verhältnisstatistiken sowie Methoden wie *qpWave* und *qpAdm* implementiert. Seit ihrer Veröffentlichung im Jahr 2012 wurde diese Software in fast allen

neueren Studien zur alten menschlichen DNA verwendet (70, 96–103), und ist zu einem Standardwerkzeug für populationsgenetische Analysen auch bei anderen Spezies geworden (104–106). Obwohl die ADMIXTOOLS-Programme leistungsstark und umfassend sind, erfordern sie einen erheblichen Anteil an manueller Arbeit, um selbst die grundlegendsten Analysen durchzuführen. Historisch gesehen war dies bisher kein kritisches Thema, da die Zahl der alten menschlichen Proben in populationsgenetischen Studien recht begrenzt war (60, 74, 107). Da die Zahl der alten Genome jedoch weiterhin rapide ansteigt (99, 103, 108), wird die Durchführung von computergestützten Analysen, die vollständig reproduzierbar sind, zu einer Herausforderung (94, 95, 109).

Da die Arbeit, die ich in *Kapitel 1* (59) beschrieben habe, die Durchführung von Tausenden von Vermischungstests an Hunderten von Proben unter Verwendung der ADMIXTOOLS-Programme umfasste, entwickelte ich eine programmierbare, einfach zu bedienende Schnittstelle zur Prüfung populationsgenetischer Hypothesen und zur Schätzung von Vermischungsparametern unter Verwendung der ADMIXTOOLS-Suite als Backend. Die Software mit dem Namen *admixr* ist als Paket für die Programmiersprache R implementiert, die ein De-facto-Standardrahmen für statistische Analysen und graphische Darstellungen ist (110). Dieses Paket ermöglicht es, alle Schritte der ADMIXTOOLS-Analysen direkt in R durchzuführen, wodurch manuelle Arbeit und fehleranfälliges Skripten eliminiert werden, und nutzt dennoch intern die effizienten und gut getesteten ADMIXTOOLS-Algorithmen. Am wichtigsten ist jedoch, dass *admixr* explizit mit Blick auf eine vollständige Automatisierung und vollständige Reproduzierbarkeit entwickelt wurde, um den Bereich der alten DNA zu reproduzierbareren Analysen zu verhelfen (94, 95, 109). Seit dem Jahr seiner Veröffentlichung (109), wurde das Paket für populationsgenetische Forschung an Menschen und anderen Spezies angenommen (111–122).

Schlussfolgerungen

Die in dieser Thesis beschriebene Arbeit hat zu mehreren Aspekten der Untersuchung der alten menschlichen Introgression beigetragen, sowohl aus demographischer Sicht als auch unter dem Gesichtspunkt der natürlichen Selektion.

Zunächst habe ich gezeigt, dass der Anteil der Neandertaler-Abstammung beim modernen Menschen seit der Introgression im Wesentlichen konstant geblieben ist, und dass dies mit den Erwartungen übereinstimmt, die von der Theorie der Populationsgenetik diktiert werden. Dadurch wurde die weithin akzeptierte Vorstellung korrigiert, dass der Anteil der Neandertaler-Abstammung bei den Eurasiern monoton von etwa 6% zur Zeit der Introgression auf etwa 1-2% bei den heutigen Menschen zurückgegangen ist (70). Ich habe gezeigt, dass die wahrscheinlichste Ursache für die falsche Schlussfolgerung des Rückgangs der Neandertaler-Abstammung in früheren Studien ein nicht berücksichtigter Genfluss

zwischen Westeurasiern und Afrikanern ist. Die neu etablierte, stabile Trajektorie der Neandertalerabstammung kann nun als Beschränkung für zukünftige Modelle von Selektion, die auf eine archaische menschliche Introgression wirkt, dienen. Sie deutet auch darauf hin, dass die bisher angenommene Verringerung der Neandertaler-Abstammung bei Europäern durch Genfluss aus einer Population mit geringer bis gar keiner Neandertaler-Introgression (53), nur einen begrenzten Effekt gehabt haben dürfte, falls sie überhaupt auftrat. Im Allgemeinen unterstreichen meine Ergebnisse die Notwendigkeit einer sorgfältigen Untersuchung der Annahmen, die hinter den Schätzungen der Populationsvermischung stehen, die über eine einfache intuitive Argumentation hinausgehen, idealerweise durch umfangreiche Simulationen, deren Durchführung selbst für Daten in der Größenordnung von ganzen Genomen heute trivial ist (123). Darüber hinaus weisen meine Ergebnisse auf einen vielversprechenden Weg für zukünftige Forschungen über den genauen Zeitpunkt, die geographische Ausdehnung und die Quelle des Genflusses von West-Eurasien nach Afrika hin, wobei die dichte räumliche und zeitliche Probenahme von alter DNA verschiedener westeurasischer Populationen genutzt wird. So hat sich beispielsweise eine kürzlich durchgeführte Studie eingehender mit den Auswirkungen dieses Genflusses befasst und gezeigt, dass er einen Großteil der offensichtlichen Unterschiede in der Abstammung der Neandertaler zwischen Europa und Ostasien erklären kann (56). Schließlich habe ich gezeigt, dass es einen unvoreingenommenen Weg gibt, die Proportionen der Neandertaler-Abstammung abzuschätzen, was besonders wichtig sein wird, da DNA von Individuen, die nahe an den Zeitpunkt der Neandertaler-Introgression heran lebten, verfügbar wird (124). Ein weiteres wichtiges Ergebnis dieser Arbeit ist die Entwicklung eines R-Pakets namens *admixr*, das die Programmierung von groß angelegten, reproduzierbaren populationsgenetischen Analysen von Tausenden von heutigen und alten Proben ermöglicht.

Zweitens habe ich die erste umfassende Analyse der Y-Chromosomen von Denisova-Menschen und Neandertalern vorgestellt. Ich habe gezeigt, dass ihre phylogenetischen Beziehungen zu den Y-Chromosomen des modernen Menschen die ihrer mitochondrialen Genome widerspiegeln, wobei die Y-Chromosomen der Neandertaler enger mit dem modernen Menschen verwandt sind als zu denen der Denisova-Menschen. Dies deutet darauf hin, dass ähnlich wie ihre mitochondriale DNA, die Y-Chromosomen der Neandertaler durch Genfluss aus einer mit den frühen modernen Menschen verwandten Population vollständig ersetzt wurden. Diese Arbeit erweitert die Sammlung an immer zahlreicher werdenden Beweisen für den alten Genfluss von modernen Menschen zu Neandertalern (85–87) und insbesondere die Sammlung an Beweisen für die Präsenz moderner Menschen außerhalb Afrikas mindestens hunderttausend Jahre vor der Hauptmigration aus Afrika (125, 126). Da alle drei Y-Chromosomen des Neandertalers, die ich analysiert habe, aus Westeurasiern stammen, wird eine umfangreichere geographische Probenahme erforderlich sein, um das Ausmaß des Austauschs des Y-Chromosoms zu

bestimmen. Angesichts der Tatsache, dass sich das Signal des mtDNA-Austauschs über den gesamten eurasischen Kontinent erstreckt (85) und angesichts der Belege für dramatische kontinentweite Umwälzungen bei Neandertalern im Allgemeinen (88, 127), kann man davon ausgehen, dass der Austausch der Neandertaler-Y-Chromosomen sogar die östlichste Ecke ihres Lebensraums in Sibirien erreicht hat. In ähnlicher Weise wird die Beprobung der Y-Chromosomen älterer Neandertaler dazu beitragen, den Zeitpunkt der Austauschdynamik zu bestimmen und die Selektionsparameter, die den Austauschprozess steuern, einzugrenzen. Insbesondere auf der Grundlage des Alters der Sima de los Huesos-Individuen und des phylogenetischen Status ihrer mtDNA (83, 84), sollten ihre Y-Chromosomen mit den Y-Chromosomen der Denisova-Menschen eine Gruppe bilden. Falls mehr Sequenzdaten der Sima de los Huesos-Homininen dies zeigen, könnten ihre Y-Chromosomen einen wichtigen Meilenstein darstellen, um festzustellen, wann der Genfluss in die Neandertaler stattgefunden hat. Obwohl meine Simulationen der Austauschdynamik der Y-Chromosomen der Neandertaler zeigen, dass ein einfaches Modell der schwächeren purifizierenden Selektion auf der Neandertaler-Linie ausreicht, um den beobachteten Austausch der Y-Chromosomen allein zu erklären, bleibt der genaue Funktionsmechanismus, der den Austausch vorantrieb, unklar. Die Faktoren, die sich negativ auf die männliche Fertilität auswirken, sind zahlreich und komplex, und viele von ihnen beinhalten strukturelle Veränderungen und chromosomale Neuarrangements (128), von denen wegen der Qualitätsgrenzen der verfügbaren alten DNA keine identifiziert werden konnten. Zukünftige technologische Fortschritte in der Extraktion alter DNA und in der Sequenzieretechnologie werden jedoch hoffentlich mehrere Y-Chromosomen mit hoher Abdeckung von Neandertalern und Denisova-Menschen hervorbringen, ohne auf die Anreicherung von DNA zurückgreifen zu müssen. In diesem Fall könnte die Neuzusammensetzung der Y-Chromosomen von Neandertalern und Denisova-Menschen und ihr Vergleich mit modernen menschlichen Y-Chromosomen bestimmen, welche Veränderungen zur geringeren Fitness der Y-Chromosomen von Neandertalern und Denisova-Menschen und folglich zu ihrer letztendlichen Ablösung geführt haben könnten.

Chapter 1

Limits of long-term selection against Neandertal introgression

Petr, M., Pääbo, S., Kelso, J. & Vernot, B.

PNAS, 116, 1639-1644 (2019)



Limits of long-term selection against Neandertal introgression

Martin Petr^a, Svante Pääbo^a, Janet Kelso^{a,1,2}, and Benjamin Vernot^{a,1,2}

^aDepartment of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Edited by Sarah A. Tishkoff, University of Pennsylvania, Philadelphia, PA, and approved December 12, 2018 (received for review August 22, 2018)

Several studies have suggested that introgressed Neandertal DNA was subjected to negative selection in modern humans. A striking observation in support of this is an apparent monotonic decline in Neandertal ancestry observed in modern humans in Europe over the past 45,000 years. Here, we show that this decline is an artifact likely caused by gene flow between modern human populations, which is not taken into account by statistics previously used to estimate Neandertal ancestry. When we apply a statistic that avoids assumptions about modern human demography by taking advantage of two high-coverage Neandertal genomes, we find no evidence for a change in Neandertal ancestry in Europe over the past 45,000 years. We use whole-genome simulations of selection and introgression to investigate a wide range of model parameters and find that negative selection is not expected to cause a significant long-term decline in genome-wide Neandertal ancestry. Nevertheless, these models recapitulate previously observed signals of selection against Neandertal alleles, in particular the depletion of Neandertal ancestry in conserved genomic regions. Surprisingly, we find that this depletion is strongest in regulatory and conserved noncoding regions and in the most conserved portion of protein-coding sequences.

Neandertal | selection | introgression | modern human | demography

Interbreeding between Neandertals and modern humans ~55,000 y ago has resulted in all present-day non-Africans inheriting at least 1–2% of their genomes from Neandertal ancestors (1, 2). There is significant heterogeneity in the distribution of this Neandertal DNA across the genomes of present-day people (3, 4), including a reduction in Neandertal alleles in conserved genomic regions (3). This has been interpreted as evidence that some Neandertal alleles were deleterious for modern humans and were subject to negative selection following introgression (3, 5). Several studies have suggested that low effective population sizes (N_e) in Neandertals led to decreased efficacy of purifying selection and the accumulation of weakly deleterious variants. Following introgression, these deleterious alleles, along with linked neutral Neandertal alleles, would have been subjected to more efficient purifying selection in the larger modern human population (6, 7).

In apparent agreement with this hypothesis, a study of Neandertal ancestry in a set of anatomically modern humans from Upper-Paleolithic Europe used two independent statistics to conclude that the amount of Neandertal DNA in modern human genomes decreased monotonically over the last 45,000 y (Fig. 1*A*, dashed line) (8). This decline was interpreted as direct evidence for continuous negative selection against Neandertal alleles in modern humans (8–11). However, it was not formally shown that selection on deleterious introgressed variants could produce a decline in Neandertal ancestry of the observed magnitude. Nevertheless, this decrease in Neandertal ancestry—together with the suggestion of a higher burden of deleterious alleles in Neandertals—are now commonly invoked to explain the fate of Neandertal ancestry in modern humans (9–12).

Here, we reexamine estimates of Neandertal ancestry in ancient and present-day modern humans, taking advantage of a second high-coverage Neandertal genome that recently became available (13). This allows us to avoid some key assumptions about modern human demography that were made in previous

studies. Our analysis shows that the Neandertal ancestry proportion in Europeans has not decreased significantly over the last 45,000 y. Using simulations of selection and introgression, we show that a model of weak selection against deleterious Neandertal variation also does not predict significant changes in Neandertal ancestry during the time period covered by existing ancient modern human samples. In contrast, these simulations do predict a depletion of Neandertal ancestry around functional genomic regions. We then use our updated Neandertal ancestry estimates to examine the genomic distribution of introgressed Neandertal DNA and find that selection against introgression was strongest in regulatory and conserved noncoding regions compared with protein-coding sequence (CDS), suggesting that regulatory differences between Neandertals and modern humans may have been more extreme than protein-coding differences.

Results

Previous Neandertal Ancestry Estimate. A number of methods have been developed to quantify Neandertal ancestry in modern human genomes (14). Among the most widely used is the f_4 -ratio statistic, which measures the fraction of drift shared with one of two parental lineages to determine the proportion of ancestry, α , contributed by that lineage (Fig. 1 and *SI Appendix, Fig. S1*) (15, 16). Although they have been used to draw inferences about gene flow between archaic and modern human populations, f_4 -ratio statistics are known to be sensitive to violations of the underlying population model (15). Estimating α , the proportion

Significance

Since the discovery that all non-Africans inherit 2% of their genomes from Neandertal ancestors, there has been a great interest in understanding the fate and effects of introgressed Neandertal DNA in modern humans. A number of recent studies have claimed that there has been continuous selection against introgressed Neandertal DNA over the last 55,000 years. Here, we show that there has been no long-term genome-wide removal of Neandertal DNA, and that the previous result was due to incorrect assumptions about gene flow between African and non-African populations. Nevertheless, selection did occur following introgression, and its effect was strongest in regulatory regions, suggesting that Neandertals may have differed from humans more in their regulatory than in their protein-coding sequences.

Author contributions: M.P., S.P., J.K., and B.V. designed research; M.P., J.K., and B.V. performed research; M.P. and B.V. contributed new reagents/analytic tools; M.P., J.K., and B.V. analyzed data; and M.P., S.P., J.K., and B.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹J.K. and B.V. contributed equally to this work.

²To whom correspondence may be addressed. Email: kelso@eva.mpg.de or benjamin_ernot@eva.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814338116/-DCSupplemental.

Published online January 15, 2019.

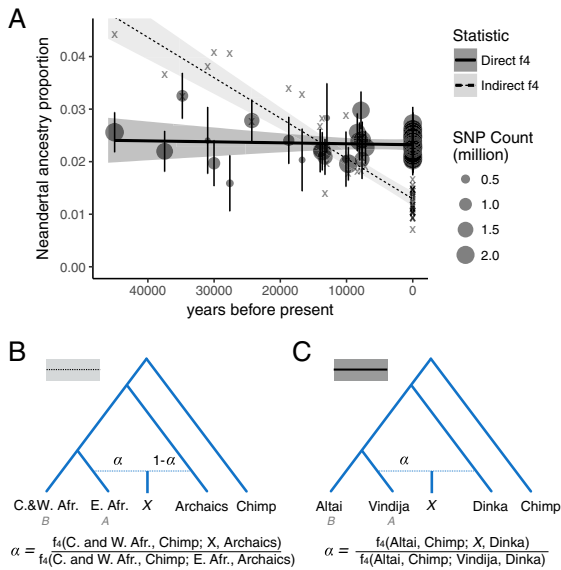


Fig. 1. Direct and indirect f_4 -ratio estimates of Neandertal ancestry. (A) Best linear fits for indirect and direct f_4 -ratio estimates of Neandertal ancestry in ancient and modern West Eurasians (solid points for direct f_4 -ratio, “x” for indirect f_4 -ratio). Shaded areas are 95% CIs (SI Appendix, section S1). (B) Tree model and formula used for the indirect f_4 -ratio. (C) Tree model and formula used for the direct f_4 -ratio. Present-day individuals are West Eurasians from the SGDP panel, excluding individuals from the Near East (Neandertal ancestry for all West Eurasians shown in SI Appendix, Fig. S7).

of ancestry in X contributed by a lineage A , requires a sister lineage B to lineage A which does not share drift with X after separation of B from A (SI Appendix, Fig. S1). Fu et al. (8) used an f_4 -ratio statistic to infer the contribution from an archaic lineage by first estimating the proportion of East African ancestry in a non-African individual X , under the assumption that Central and West Africans (B) are an outgroup to the East African lineage (A) and to the modern human ancestry in non-Africans. Defining this East African ancestry proportion as $\alpha = f_4(\text{C. and W. Africans, Chimp; X, Archaics})/f_4(\text{C. and W. Africans, Chimp; E. Africans, Archaics})$, the proportion of archaic ancestry was then calculated simply as $1 - \alpha$, under the assumption that all ancestry that is not of East African origin must come from an archaic lineage (8). We refer to this statistic as an “indirect f_4 -ratio.”

Given the sensitivity of the f_4 -ratio method to violations of the underlying population models (15), we explored the validity of assumptions on which this calculation was based. In addition to the topology of the demographic tree, which has recently been shown to be incorrect (17), the indirect f_4 -ratio assumes that the relationship between Africans and West Eurasians has remained constant over time (8). However, our understanding of modern human history and demography have been challenged by new fossil discoveries (18) and the analysis of ancient DNA, with several studies documenting previously unknown migration events in both West Eurasia (19) and Africa (17, 20, 21). Furthermore, an f_4 statistic sensitive to changes in the relationships between West Eurasians and various African populations [formulated as $f_4(\text{Ust’-Ishim, X; African, Chimp})$, where X is a West Eurasian individual] shows increasing allele sharing between West Eurasians and Africans over time (SI Appendix, Fig. S24). In contrast, $f_4(\text{Ust’-Ishim, Papuan; African, Chimp})$ is not significantly different from zero ($|Z| < 1$ when using Dinka, Yoruba,

or Mbuti in the third position of the f_4 statistic), demonstrating that this trend is not shared by all non-Africans.

To evaluate the sensitivity of the indirect f_4 -ratio to migration events, we performed neutral simulations of Neandertal, West Eurasian, and African demographic histories (Fig. 2). All simulations included introgression from Neandertals into West Eurasians, and varying levels of migration between Africans and West Eurasians, and between African populations. We find that gene flow from West Eurasians into Africans leads to misestimates of Neandertal ancestry when using the indirect f_4 -ratio statistic, and results in the incorrect inference of a continuous decline in Neandertal ancestry. This decline is not observed in the true simulated Neandertal ancestry (Fig. 2A). The magnitude of this bias depends on the total amount of West Eurasian gene flow into Africa, with larger amounts leading to apparent steeper declines (Fig. 2A). Additionally, gene flow between the two African populations used in the indirect f_4 -ratio calculation leads to overestimation of the true level of Neandertal ancestry (Fig. 2C). Overall, we find that a combination of West Eurasian migration to Africa and gene flow between African populations can produce patterns that are very similar to those observed in the empirical data (Fig. 2D and SI Appendix, Fig. S34). However, we caution that effective population sizes and the timing of migration also affect these estimates (SI Appendix, Fig. S3), and that there are likely many additional models that match the empirical data.

We note that an independent statistic, using a different set of genomic sites in the same ancient individuals, had been used as a second line of evidence for an ongoing decrease in Neandertal ancestry (8). This statistic, which we refer to as the “admixture array statistic,” measures the proportion of Neandertal-like alleles in a given sample at sites where present-day Yoruba individuals carry a nearly fixed allele that differs from homozygous sites in the Altai Neandertal (22). Much like the indirect f_4 -ratio, we find that the admixture array statistic is affected by gene flow from non-Africans into Africans and incorrectly infers a decline in the Neandertal ancestry over time (Fig. 2D).

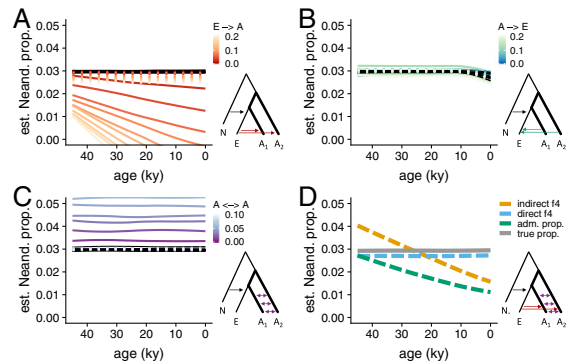


Fig. 2. Neandertal ancestry estimates in neutral simulations of migration. Genomic data were simulated under a base model of 3% Neandertal admixture, $N_e = 6,000$ in Europeans and $N_e = 14,000$ in two African populations (SI Appendix, Fig. S8, section S2). (A–C) The effect of three migration parameters on direct and indirect f_4 -ratio estimates of Neandertal ancestry (dotted and solid colored lines, respectively). “Total migration” is shown, that is, gm , where g is generations of migration, and m is the proportion of the target population composed of migrants in each generation. If present, continuous migration between A_1 and A_2 begins 40 kya and migration between Europe and Africa begins 5 kya. True Neandertal ancestry proportions are shown in black, and closely match the direct f_4 -ratio estimates (mean absolute difference from truth for indirect f_4 -ratio is 2.6%, 0.12%, and 2.8% for A, B, and C respectively; for direct f_4 -ratio 0.25%, 0.05%, and 0.06%). (D) Simulations of an example demographic model with migration parameters 0.09, 0.0, and 0.1 for $E \rightarrow A$, $A \rightarrow E$, and $A \leftrightarrow A$, respectively, which approximate the empirical direct and indirect f_4 -ratios (Fig. 1A).

Given the indirect f_4 -ratio's sensitivity to modern human demography, combined with our incomplete understanding of human migrations, we sought to reevaluate the patterns of Neandertal ancestry in modern humans in a more robust manner.

A Robust Statistic to Estimate Neandertal Ancestry. The recent availability of a second high-coverage Neandertal genome allows us to estimate Neandertal ancestry using two Neandertals—an individual from the Altai Mountains, the so-called “Altai Neandertal” (23) and an individual from the Vindija Cave in Croatia, the so-called “Vindija Neandertal” (13). Specifically, we can estimate the proportion of ancestry coming from the Vindija lineage into a modern human (X) using the Altai Neandertal as a second Neandertal in an f_4 -ratio calculated as $f_4(\text{Altai, Chimp}; X, \text{African})/f_4(\text{Altai, Chimp}; \text{Vindija, African})$, which we refer to as a “direct f_4 -ratio” (Fig. 1C and *SI Appendix, Fig. S1*). Note that unlike the indirect f_4 -ratio described previously, the f_4 -ratio in this formulation does not make assumptions about deep relationships between modern human populations (Fig. 1C and *SI Appendix, Fig. S1*). Instead, it assumes that any Neandertal population that contributed ancestry to X formed a clade with the Vindija Neandertal. Recent analyses showed that this is the case for all non-African populations studied to date, including the ancient modern humans in this study (13, 24). When calculated on the simulations described above, we find that the direct f_4 -ratio is more robust than the indirect f_4 -ratio (Fig. 2). In fact, its temporal trajectory always closely matches the true simulated Neandertal ancestry trajectory, regardless of the specific parameters of gene flow between non-Africans and Africans (Fig. 2). We note that gene flow from West Eurasians into Africans, which introduces introgressed Neandertal alleles into Africa, produces a slight underestimate of Neandertal ancestry in all samples (Fig. 24). This is in agreement with empirical direct f_4 -ratio estimates, which vary depending on the African population used in the calculation, with African populations known to carry West Eurasian ancestry (e.g., Mozabite, Saharawi) (17, 25) generating the lowest estimates (*SI Appendix, Fig. S4*). Crucially, when we use the direct f_4 -ratio to estimate the trajectory of Neandertal ancestry in ancient and present-day Europeans, we observe nearly constant levels of Neandertal ancestry over time (Fig. 1A, points and solid line) and find that a null model of zero slope can no longer be rejected (Fig. 1A, $P = 0.36$, estimated via resampling as described in *SI Appendix, section S1*).

We note that these estimates are based on a relatively small number of individuals, especially for older time points, and that the CIs are wide. For example, we cannot reject a linear decline in Neandertal ancestry of approximately half a percent over the timespan of this dataset (95% CI -0.51 – -0.37%). Additionally, these analyses are performed on SNPs that were ascertained largely in present day individuals. To examine the effects of such ascertainment, we split the dataset based on the ascertainment used and recalculated the direct and indirect f_4 -ratios on each of the subsets (*SI Appendix, Fig. S5*). Although the slopes show some variability, in all but one ascertainment subset the direct f_4 -ratio cannot reject a slope of 0, whereas the indirect f_4 -ratio consistently rejects a slope of 0, suggesting that these results are robust to the effects of ascertainment (*SI Appendix, Fig. S5*). In addition to calculating direct f_4 -ratio estimates, we estimated Neandertal ancestry proportions using the qpAdm method (26) and obtained similar results (null model of zero slope using Neandertal ancestry point estimates cannot be rejected with $P = 0.17$).

Our observation that there has been no change in Neandertal ancestry over the past 45,000 y has several implications for our understanding of the fate of Neandertal DNA in modern humans. First, it constrains the timescale during which selection could have significantly affected the average genome-wide Neandertal ancestry in modern humans, an issue addressed below in more detail. Second, a previous analysis of a 40 ky old individual (“Tianyuan”) from East Asia applied the indirect f_4 -ratio statistic to estimate his

Neandertal ancestry proportion at 5% (27). When we apply the direct f_4 -ratio statistic for this individual, we arrive at a value of $\sim 2.1\%$ (using Dinka as the African group in the calculation). Third, it has consequences for the so-called “dilution” hypothesis, which suggests that lower levels of Neandertal ancestry in Europeans compared with East Asians can be explained by dilution of Neandertal ancestry in Europeans due to admixture with a hypothetical Basal Eurasian population that carried little to no Neandertal ancestry (19, 28). Previous studies have found Basal Eurasian ancestry in all modern and some ancient Europeans [in this study, four ancient individuals show evidence of Basal Eurasian ancestry: Satsurblia (15 kya), Kotias (10 kya), Ranchot88 (10 kya), and Stuttgart (8 kya), *SI Appendix, Fig. S6*] (8, 19). Our finding that there is no ongoing decline in Neandertal ancestry in Europeans suggests that Neandertal ancestry in Europe has not been diluted in a significant way by gene flow from Basal Eurasians. Specifically, we find no difference in Neandertal ancestry in European individuals with and without Basal Eurasian ancestry (direct f_4 -ratio mean 2.31% vs. 2.38%, respectively; $P = 0.36$). However, given the small number of relevant samples we also cannot exclude that there could be up to 13% less Neandertal ancestry in individuals with Basal Eurasian ancestry, or as much as 6% more Neandertal ancestry in individuals without Basal Eurasian ancestry (95% CI).

In contrast, we do find that present-day Near Easterners carry significantly less Neandertal ancestry than Europeans (direct f_4 -ratio mean 2.03% vs. 2.33%; $P = 0.001$; *SI Appendix, Fig. S7A*). Furthermore, present-day populations in the Near East show even stronger signals of admixture with a deeply divergent modern human lineage than observed in the rest of West Eurasians (*SI Appendix, Fig. S7B*), suggesting that they carry additional ancestry components that are not present in Europe and that could potentially contribute to lower Neandertal ancestry in the Near East. We note, however, that a simple model of admixture from Africa into Near East would be expected to produce a similar f_4 statistics difference between Near East and the rest of West Eurasia and could also explain lower values of Neandertal ancestry in this population.

Long-Term Dynamics of Selection Against Introgressed DNA. Our observation that Neandertal ancestry levels did not significantly decrease from $\sim 45,000$ y ago until today is seemingly at odds with the hypothesis that lower effective population sizes in Neandertals led to an accumulation of deleterious alleles, which were then subjected to negative selection in modern humans (3, 8–10). To investigate the expected long-term dynamics of selection against Neandertal introgression under this hypothesis, we simulated a model of the human genome with empirical distributions of functional regions and selection coefficients, extending a strategy previously applied by Harris and Nielsen (6). We simulated modern human and Neandertal demography, including a low long-term effective population size (N_e) in Neandertals (Neandertal $N_e = 1,000$ vs. modern human $N_e = 10,000$) and 10% introgression at 55 kya (2,200 generations ago, assuming generation time of 25 y). To track the changes in Neandertal ancestry following introgression, we placed fixed Neandertal–human differences as neutral markers, both outside regions that accumulated deleterious mutations (to study the effect of negative selection on linked genome-wide neutral Neandertal variation) as well as within regions directly under selection (to track the effect of negative selection itself) (Fig. 3A).

Similar to Harris and Nielsen (6), we observed abrupt removal of Neandertal alleles from the modern human population during the first ~ 10 generations after introgression, followed by quick stabilization of Neandertal ancestry levels (Fig. 3B). Compared with empirical estimates of Neandertal ancestry, we find a better fit between these simulations and the direct f_4 -ratio estimate than with the indirect f_4 -ratio estimate, suggesting that our direct Neandertal ancestry estimates are consistent with theoretical expectations of genome-wide selection against introgression

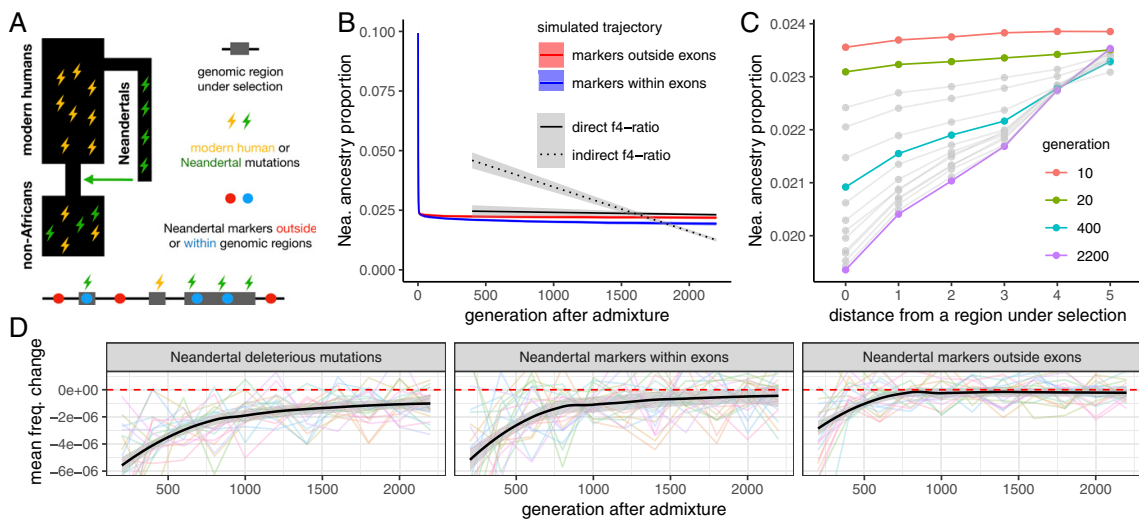


Fig. 3. Simulations of selection against Neandertal ancestry. (A) Deleterious mutations (lightning bolts) accumulate in realistically distributed exonic sequence in modern humans and Neandertals. These regions accumulate additive, deleterious mutations, using a mutation rate of 10^{-8} per base pair per generation. To track the dynamics of Neandertal ancestry over time, neutral Neandertal markers are placed within (blue dots) and between (red dots) exons on all Neandertal chromosomes before introgression. (B) Simulated Neandertal ancestry proportions across 55 ky, in exonic and nonexonic sequence, averaged over 20 simulation replicates. Empirical observations from Fig. 1A are shown for comparison. Initial introgression levels were simulated at 10%. (C) Depletion of simulated Neandertal ancestry at neutral markers over time as a function of distance to regions under selection. Markers in bin 0 are those falling within exons; bins 1–5 represent quintiles of distance to the nearest exon. (D) Changes in frequencies of neutral Neandertal markers and deleterious Neandertal mutations over time, starting from generation 200. Each line shows average allele frequency changes over one simulation replicate. Black lines show smooth fits of these averages over 20 replicates.

(Fig. 3B). Specifically, simulations show -0.004% change in Neandertal ancestry over 45 ky; in the empirical data this slope is not rejected using the direct f_4 -ratio ($P = 0.29$), but is significantly different from the indirect f_4 -ratio ($P < 0.001$).

Because many factors can potentially influence the efficacy of negative selection, and no model fully captures all of these, we next sought to determine whether there is a combination of model parameters that could potentially lead to long-term continuous removal of Neandertal ancestry over time. Surprisingly, we failed to find a model which would produce a significant decline over time, although we tried by: (i) decreasing the long-term Neandertal N_e before introgression (making purifying selection in Neandertals even less efficient), (ii) increasing the N_e of modern humans after introgression (i.e., increasing the efficacy of selection against introgressed alleles), (iii) artificially increasing the deleteriousness of Neandertal variants after introgression (approximating a “hybrid incompatibility” scenario), (iv) simulating mixtures of dominance coefficients, or by (v) increasing the total amount of functional sequence (thereby increasing the number of accumulated deleterious variants in Neandertals and modern humans) (SI Appendix, Figs. S9–S13). Varying these factors primarily affected the magnitude of the initial removal of introgressed DNA by increasing the number of perfectly linked deleterious mutations in early Neandertal–modern human offspring (decreasing their fitness compared with individuals with less Neandertal ancestry), which in turn influenced the final level of Neandertal ancestry in the population (SI Appendix, Figs. S9–S13).

The depletion of Neandertal ancestry around functional genomic elements in modern human genomes has also been taken as evidence for selection against Neandertal introgressed DNA (3, 8). We next examined the genomic distribution of Neandertal markers at different time points in our simulations to determine whether our models can recapitulate these signals. In agreement with empirical results in present-day humans (3), we found a strong negative correlation between the proportion of Neandertal

introgression surviving at a locus and distance to the nearest region under selection (Fig. 3C). Furthermore, we found that the strength of this correlation increases over time, with the bulk of these changes occurring between 10 and 400 generations postadmixture [mean Pearson’s correlation coefficient $\rho = 0.07, 0.79, 0.96$ at generations 10, 400, and 2,200, respectively (SI Appendix, Fig. S15)]. We note that this time period predates all existing ancient modern human sequences, frustrating any current comparison with empirical data. However, despite no apparent change in genome-wide Neandertal ancestry proportion over time, we observe a smaller though still significant decrease in linked Neandertal ancestry during the time period for which modern human sequences exist (~400–2,200 generations post-admixture) (Fig. 3C and B). Indeed, by looking at the average per-generation changes in frequencies of simulated Neandertal mutations (that is, derivatives of allele frequencies in each generation), we observe the impact of negative selection on linked neutral Neandertal markers until at least ~700 generations post admixture (Fig. 3D) and find that it closely follows the pattern of introgressed deleterious mutations (Fig. 3D). After this period of gradual removal, selection against linked neutral variation slows down significantly as genome-wide Neandertal ancestry becomes largely unlinked from regions that are under negative selection (Fig. 3D). In contrast, the selected variants themselves are still removed, although at increasingly slower rates (Fig. 3D). Due to this slow rate, and the small contribution these alleles make to genome-wide Neandertal ancestry, their continued removal has little impact on the slope of Neandertal ancestry over time.

Neandertal DNA Is Depleted in Regulatory and Conserved Noncoding Sequence. We next sought to leverage the direct f_4 -ratio in analyses of selection against introgression in functional genomic regions. Although previous studies have identified a depletion of Neandertal DNA in genomic regions with a high degree of evolutionary conservation, these studies have relied on maps of introgressed haplotypes (3, 29). Such maps may lack power to detect introgressed

Neandertal DNA in highly conserved regions, as these regions may contain fewer informative sites carrying Neandertal–modern human differences. Furthermore, previous studies of negative selection against introgressed Neandertal DNA divided the genome into bins based on measures of evolutionary conservation, such as B values (30), which are not easily interpreted in terms of functional significance. To determine whether particular functional classes of genomic sites are differently affected by Neandertal introgression, we partitioned the human genome by functional annotation obtained from Ensembl v91 (31), and by primate conserved regions inferred using phastCons (32). For each annotation category, we estimated the Neandertal ancestry proportion in non-African Simons Genome Diversity Project (SGDP) individuals (excluding Oceanians) using the direct f_4 -ratio (Fig. 4).

In seeming contrast with previous studies (3, 8), we observed no significant depletion of Neandertal ancestry in CDS compared with intronic and intergenic regions (referred to as “gap” regions below) (average direct f_4 -ratio $\sim 1.94\%$ in both; Fig. 4). However, we did identify a striking depletion of Neandertal ancestry in both promoters and phastCons conserved regions (1.15% and 0.95%), with both containing significantly less Neandertal ancestry than gap regions ($P = 0.004$ and $P < 0.0001$, estimated via resampling as described in *SI Appendix*, section S1). We note that 62% of CDS overlaps with phastCons regions (21% of phastCons conserved tracks overlap CDS); indeed, conserved CDS has a lower Neandertal ancestry estimate (1.25%) than overall CDS, although not as low as all phastCons regions (Fig. 4). These results suggest that previously observed depletions in conserved and genic regions may not have been driven primarily by protein-coding differences between Neandertals and modern humans, as was previously assumed, but rather by differences in promoters and other noncoding conserved sequence. This hypothesis is supported by several recent studies of the effects of introgressed Neandertal sequences, including those with signatures of adaptive introgression, which found that surviving functional introgressed haplotypes have their major influence on gene expression regulation (33–37).

We note that the lack of a depletion in CDS does not fit the observations from our simulations (Fig. 3C). Assuming additivity,

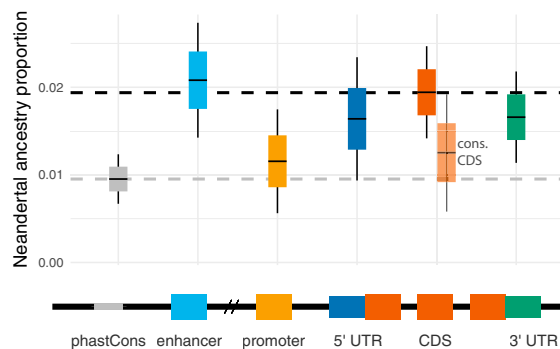


Fig. 4. Neandertal ancestry estimates by genomic region. (Top) Direct f_4 -ratio estimates of Neandertal ancestry in all non-African SGDP individuals except Oceanians (known to carry Denisovan ancestry in addition to Neandertal ancestry) (25), with SNPs partitioned by functional annotation (Ensembl) or conservation (phastCons); “gap” combines intronic and intergenic sequence (dashed black line). Many annotation categories overlap other categories (*SI Appendix*, Table S1)—the largest is the 62% of protein-coding sequence which overlaps phastCons conserved elements (translucent orange). To minimize the noise in Neandertal ancestry estimates for small subsets of the genome, we calculated the direct f_4 -ratio using all SGDP Africans, except those that carry a high proportion of Neandertal alleles (Mozabite, Saharawi, Ju’hoan North, Khomani San and Somali in *SI Appendix*, Fig. S4). Gray dashed line shows mean Neandertal ancestry in conserved phastCons regions. (Bottom) Idealized representation of genomic regions.

Petr et al.

and a distribution of fitness effects (DFEs) derived from the frequency spectra of mutations altering coding sequence (38), these simulations predict a reduction of 5–17% Neandertal ancestry versus nonselected regions, depending on distance from selected regions (Fig. 3C). In addition, the reduction in simulations is much smaller than the empirical depletions of promoter and phastCons regions (40% and 51%, respectively). Together, these demonstrate that the actions of selection against Neandertal sequence are not fully captured by the models presented here. Although it is beyond the scope of this work, it may be possible to leverage distributions of Neandertal ancestry in studying the action of selection in noncoding sequence. Challenges associated with such work include the uncertainty of the DFE of mutations affecting noncoding sequence, and their dominance coefficients, potential epistatic effects of regulatory mutations, as well as the fact that a single deleterious mutation can affect a region falling into multiple functional categories at once (*SI Appendix*, Table S1).

Conclusions

Our reevaluation of Neandertal ancestry in modern human genomes indicates that overall levels of Neandertal ancestry in Europe have not significantly decreased over the past 45,000 y, and that previous observations of continuous Neandertal ancestry decline were likely an artifact of unaccounted-for gene flow increasing allele sharing between West Eurasian and African populations. Nevertheless, we do find evidence of selection against Neandertal DNA in the genome-wide distribution of Neandertal ancestry, with such ancestry depleted in promoter and other noncoding conserved DNA more strongly than in protein-coding sequence, raising the possibility that Neandertals may have differed more from modern humans in their regulatory variants than in their protein-coding sequences, and that regulatory variation may provide a richer template for selection to act upon.

Furthermore, simulations suggest that negative selection against introgression is expected to have the strongest impact on genome-wide Neandertal ancestry during the first few hundred generations, before the time frame for which ancient samples are currently available. The genomes of early modern humans living 55–50 kya, although difficult to obtain, may shed additional light on the process of selection against Neandertal DNA, as well as on early out-of-Africa demography.

Our findings can be extrapolated to other cases where one species or population contributes a fraction of ancestry to another species or population, a frequent occurrence in nature (5, 29, 39–41). Even in cases where the introgressing population carries a high burden of deleterious mutations, negative selection is not expected to result in an extended decrease in the overall genome-wide ancestry contributed by that population. Therefore, any long-term shifts in overall ancestry proportions over time are likely to be the result of forces other than negative selection, for example admixture with one or more other populations.

Materials and Methods

Source Code and Jupyter Notebooks. Complete source code for data processing and simulation pipelines, as well as R and Python Jupyter notebooks with all analyses, can be downloaded from the project repository on GitHub: <https://www.github.com/bodkan/nea-over-time>.

Data Processing. SNP data captured at ~ 2.2 million loci from a set of Upper Paleolithic individuals published by Fu et al. (8) were obtained from the David Reich laboratory (<https://reich.hms.harvard.edu/datasets>), and merged with previously published genotypes for the Altai Neandertal (23), Vindija Neandertal (13), Denisovan (42), and SGDP (25) to create a single EIGENSTRAT dataset. For all analyses, individuals with at least 200,000 captured sites were analyzed. SNP data captured using the “archaic admixture array” (SNP panel 4 in ref. 22) published by Fu et al. (8) were also downloaded from the Reich laboratory website and filtered to sites homozygous in the Altai and Vindija Neandertal genomes, resulting in a set of $\sim 480,000$ sites carrying nearly fixed Yoruba–Neandertal differences.

Admixture Statistics. All f_4 statistics, f_4 -ratio, and qpAdm statistics were calculated on the merged 2.2 million loci EIGENSTRAT dataset using our R package admixr (available from <https://www.github.com/bodkan/admixr>) which utilizes the ADMIXTOOLS software suite for all underlying calculations (15).

Estimates of Neanderthal Ancestry. Indirect f_4 -ratio estimates (Fig. 1A, dashed line) were calculated as $1 - f_4(\text{West and Central Africans, Chimpanzee}; X, \text{Archaics})/f_4(\text{West and Central Africans, Chimpanzee}; \text{East African, Archaics})$, where West and Central Africans are Yoruba, Mbuti, and Mende from the SGDP panel, East Africans are SGDP Dinka, and archaics are the Altai Neanderthal (23) and Denisovan (42) individuals (SI Appendix, Fig. S1), as described in the original Fu et al. study (8). Direct f_4 -ratio estimates (Fig. 1A, solid line) were calculated as $f_4(\text{Altai, Chimpanzee}; X, \text{African})/f_4(\text{Altai, Chimpanzee}; \text{Vindija, African})$ (SI Appendix, Fig. S1). Neanderthal ancestry proportions using qpAdm were estimated assuming a two-source model, with the Vindija Neanderthal and Mbuti as potential sources, and Chimpanzee, the Altai Neanderthal, and the Denisovan as outgroups. Admixture array-based Neanderthal ancestry estimates were calculated as the proportion of alleles in a test individual matching the allele seen in Neanderthals. Confidence intervals and P values were calculated using a resampling strategy described in SI Appendix, section S1.

Affinity of Ancient and Present-Day Individuals Toward Africans over Time. We calculated f_4 statistics in the form $f_4(\text{Ust'-Ishim, X}; Y, \text{Chimpanzee})$, which test for changes in the sharing of derived alleles between a series of West Eurasians (X) and population Y with respect to Ust'-Ishim, an ancient hunter-gatherer that predates the split of West and East Eurasians (43) (SI Appendix, Fig. S2). Admixture between X and Y or populations related to X and Y is expected to lead to an increase in the proportion of shared derived alleles.

Testing for the Presence of Basal Eurasian Ancestry. We used the statistic $f_4(\text{West Eurasian } W, \text{Han}; \text{Ust'-Ishim, Chimpanzee})$ to look for evidence of Basal Eurasian ancestry in a West Eurasian W (SI Appendix, Fig. S4) (28). This statistic tests if the data are consistent with a tree in which W and Han lineages form a

clade, which results in f_4 statistic not significantly different from 0. Significantly negative values are evidence for an affinity between the Ust'-Ishim and Han lineages, which could be explained by W carrying ancestry from a population that diverged from the non-African lineage before the split of Ust'-Ishim.

Neutral Coalescent Simulations. To study the effects of gene flow between non-African and African populations on various admixture statistics, we simulated different scenarios of such gene flow using a neutral coalescent programming library, msprime (44) (SI Appendix, Fig. S8). Depending on the particular analysis (Fig. 2 and SI Appendix, Fig. S2 and S3), we calculated admixture statistics (f_4 , f_4 -ratio, and admixture array proportions) as described above using SNPs extracted from each simulation run. Detailed description of the simulations can be found in SI Appendix, section S2.

Simulations of Selection. To study the dynamics of selection against Neanderthal introgression over time, we used the simulation framework SLiM 2 (45) to build a realistic model of the human genome with empirical distributions of functional and conserved regions and selection coefficients, extending and generalizing a strategy previously applied by Harris and Nielsen (6) (Fig. 3A). First, we simulated a demography of modern humans and Neanderthals (low long-term M_e) before the introgression, and let the simulated genomes accumulate deleterious mutations. Then we simulated a single pulse of admixture from Neanderthals into the non-African population at a rate of 10% and tracked the changes in Neanderthal ancestry in an admixed population at fixed neutral Neanderthal markers distributed along each Neanderthal genome before the introgression. A detailed description of our simulations and analyses of simulated data can be found in SI Appendixes, sections S3 and S4.

ACKNOWLEDGMENTS. We thank Montgomery Slatkin, Benjamin Peter, Fabrizio Mafessoni, Iosif Lazaridis, Mark Lipson, and David Reich for helpful discussions and comments on the manuscript; and Steffi Grote for processing of archaic human and SGDP datasets. The study was funded by the Max Planck Society and European Research Council Grant Agreement 694707 (to S.P.).

- Green RE, et al. (2010) A draft sequence of the Neanderthal genome. *Science* 328: 710–722.
- Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neanderthals and modern humans. *PLoS Genet* 8:e1002947.
- Sankararaman S, et al. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507:354–357.
- Vernot B, Akey JM (2014) Resurrecting surviving Neanderthal lineages from modern human genomes. *Science* 343:1017–1021.
- Schumer M, et al. (2018) Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* 360:656–660.
- Harris K, Nielsen R (2016) The genetic cost of neanderthal introgression. *Genetics* 203: 881–891.
- Juric I, Aeschbacher S, Coop G (2016) The strength of selection against neanderthal introgression. *PLoS Genet* 12:e1006340.
- Fu Q, et al. (2016) The genetic history of ice age Europe. *Nature* 534:200–205.
- Harris K, Nielsen R (2017) Q&A: Where did the Neanderthals go? *BMC Biol* 15:73.
- Yang MA, Fu Q (2018) Insights into modern human prehistory using ancient genomes. *Trends Genet* 34:184–196.
- Reich D (2018) *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (Pantheon, New York).
- Steinrücken M, Spence JP, Kamm JA, Wieczorek E, Song YS (2018) Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol Ecol* 27:3873–3888.
- Prüfer K, et al. (2017) A high-coverage Neanderthal genome from Vindija Cave in Croatia. *Science* 358:655–658.
- Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* 16:359–371.
- Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192:1065–1093.
- Peter BM (2016) Admixture, population structure and F -statistics. *Genetics* 202: 1485–1501.
- Schlebusch CM, et al. (2017) Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* 358:652–655.
- Hublin J-J, et al. (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature* 546:289–292.
- Lazaridis I, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.
- Skoglund P, et al. (2017) Reconstructing prehistoric african population structure. *Cell* 171:59–71.e21.
- van de Loosdrecht M, et al. (2018) Pleistocene North African genomes link near Eastern and sub-saharan African human populations. *Science* 360:548–552.
- Fu Q, et al. (2015) An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524:216–219.
- Prüfer K, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.
- Hajdinjak M, et al. (2018) Reconstructing the genetic history of late Neanderthals. *Nature* 555:652–656.
- Mallik S, et al. (2016) The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.
- Haak W, et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211.
- Yang MA, et al. (2017) 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr Biol* 27:3202–3208.e9.
- Lazaridis I, et al. (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature* 536:419–424.
- Sankararaman S, Mallik S, Patterson N, Reich D (2016) The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr Biol* 26:1241–1247.
- McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5:e1000471.
- Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR (2015) The Ensembl regulatory build. *Genome Biol* 16:56.
- Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
- Gittelman RM, et al. (2016) Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Curr Biol* 26:3375–3382.
- Dannemann M, Andrés AM, Kelso J (2016) Introgression of neanderthal- and denisovan-like haplotypes contributes to adaptive variation in human toll-like receptors. *Am J Hum Genet* 98:22–33.
- McCoy RC, Wakefield J, Akey JM (2017) Impacts of neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* 168:916–927.e12.
- Simonti CN, et al. (2016) The phenotypic legacy of admixture between modern humans and Neanderthals. *Science* 351:737–741.
- Dannemann M, Prüfer K, Kelso J (2017) Functional implications of Neanderthal introgression in modern humans. *Genome Biol* 18:61.
- Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.
- Jacobsen F, Omland KE (2011) Increasing evidence of the role of gene flow in animal evolution: Hybrid speciation in the yellow-rumped warbler complex. *Mol Ecol* 20:2236–2239.
- Cui R, et al. (2013) Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. *Evolution* 67:2166–2179.
- Schrider DR, Ayroles J, Matute DR, Kern AD (2018) Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genet* 14:e1007341.
- Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226.
- Fu Q, et al. (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445–449.
- Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol* 12:e1004842.
- Haller BC, Messer PW (2017) SLiM 2: Flexible, interactive forward genetic simulations. *Mol Biol Evol* 34:230–240.



Supplementary Information for

Limits of long-term selection against Neandertal introgression

Martin Petr¹, Svante Pääbo¹, Janet Kelso^{1*}, Benjamin Vernot^{1*}

¹ Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103, Leipzig, Germany.

* Contributed equally / corresponding author

Email: kelso@eva.mpg.de benjamin_vernot@eva.mpg.de

This PDF file includes:

Supplementary Information Text
Figs. S1 to S15
Table S1
References for SI reference citations

Supplementary Information Text

S1. Neandertal ancestry estimate confidence intervals and p -values

Confidence intervals on the slope of time vs Neandertal ancestry proportion were calculated empirically via resampling. For each individual, we sampled 10,000 Neandertal ancestry estimates from a normal distribution centered on the true estimate, with standard deviation equal to the standard error provided by ADMIXTOOLS. We then fit 10,000 linear models, extracted the 95% confidence intervals across all 10,000 resulting slopes. From these 10,000 slopes we can also calculate an empirical p -value for any given slope (generally for a slope of 0, but -0.004 for the comparison of simulated data vs the direct and indirect f_d -ratio estimates). For comparisons of the ratio of Neandertal ancestry between populations (e.g., individuals with and without Basal Eurasian ancestry), we similarly sampled 10,000 Neandertal ancestry estimates, and calculated an empirical p -value for a ratio of 1 between the two groups. For a comparison of functional annotation categories, we similarly resampled 10,000 Neandertal estimates from a given category, and calculated an empirical p -value that these resamplings reject the Neandertal ancestry proportion calculated for gap regions.

S2. Simulations of gene flow between non-Africans and Africans

We simulated different scenarios of gene flow between Africans and non-Africans after Neandertal introgression using the neutral coalescent programming library *msprime* (1) (Fig. S8). We used the following demographic parameters: split of a chimpanzee lineage at 6 million years ago, split of Neandertals from anatomically modern humans at 500 kya, split within Africa at 150 kya, and split of non-Africans from one of the two African lineages at 60 kya with a 5 ky long bottleneck of $N_e = 2000$. We simulated a single 3% pulse of Neandertal admixture into a constant-size non-African population at 55 kya. We sampled one chimpanzee chromosome, 4 Neandertal chromosomes sampled at 80 kya, single chromosomes from the non-African lineage sampled at regular time intervals over the time range of Upper-Paleolithic and present-day individuals from our data, and two pairs of chromosomes from the two present-day African populations. We simulated 500 replicates of 100 Mb chromosomes (Fig. S2) or 600 replicates of 0.25 Mb chromosomes

(Fig. 2, S3) using a mutation rate of 1×10^{-8} mutations per bp per generation and a recombination rate of 1×10^{-8} crossovers per bp per generation, and converted them into tables of all simulated SNPs for easier calculation of admixture statistics. For analysis of the “admixture array”, we also generated a second set of SNPs by filtering only for sites carrying fixed African-Neandertal differences (to approximate the ascertainment of the archaic admixture array – SNP panel 4 in (2)). To estimate the true Neandertal ancestry levels we examined the origin of each simulated mutation in *msprime* and extracted only those SNPs that truly originated in the Neandertal population. Using this set of sites avoids any issues caused by introduction of Neandertal alleles into Africans via gene-flow from admixed non-Africans.

In Fig. 2 and Fig. S3, we evaluated the behavior of the admixture array ancestry proportion and direct and indirect f_4 -ratio estimates under three scenarios: (i) no gene flow between Africans and non-Africans post Neandertal admixture, (ii) gene flow from non-Africans into both African populations, (iii) gene flow from one African population into non-Africans, (iv) bi-directional gene flow between Africans and non-Africans.

Using the simulated SNP sets (all SNPs and archaic admixture array-like set), we calculated direct and indirect f_4 -ratio estimates, as well as admixture array proportion estimates, as described above. Unbiased levels of Neandertal ancestry were calculated on the set of true Neandertal-derived SNPs. As the statistics can be relatively noisy, we calculated average values of each individual statistic over all simulation replicates.

S3. Simulations of selection

We used the simulation framework *SLiM* 2 (3) to build a realistic model of the human genome with empirical distributions of functional regions and selection coefficients, extending and generalizing a strategy previously applied by Harris and Nielsen (4). To obtain the coordinates of regions under negative selection, we downloaded the positions of different classes of annotated genomic regions from the Ensembl database (5) (Ensembl Genes 91 and Ensembl Regulation 91) and conserved regions from the phastConsElements46wayPrimates track from the UCSC Genome Browser (6, 7) (updated 2009-11-21). In each simulation, we encoded those regions in a genomic structure in *SLiM*'s *Eidos* programming language, maintaining the distances between

them. In order to model the heterogeneity of recombination rate along a genome in our simulations, we used empirically estimated genetic distances between all simulated genomic features using a recombination map inferred by the HapMap project (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/) (8). To approximate a mixture of strongly, weakly and nearly-neutral deleterious mutations, we used a distribution of fitness effects (DFE) estimated from the frequency spectrum of human non-synonymous mutations (9). The rate of accumulation of new mutations was set to 1×10^{-8} per bp per generation.

The simulations themselves were performed in two steps (Fig. 3A), using a combination of human and Neandertal demographic models used in previous introgression studies (4, 10). In the first step, we simulated a simplified demography of modern humans and Neandertals prior to the introgression, starting with a burn-in period of 70,000 generations, to let the simulated genomes with mutations reach an equilibrium state (the length of this burn-in period was determined as $7 * \text{ancestral human } N_e$, which was therefore set to a constant 10,000). The split of Neandertals and modern humans was set to 500,000 years ago, with N_e of Neandertals and modern humans set to constant values of 1,000 and 10,000, respectively. This burn-in period was performed for each specific simulation scenario separately. At the end of the burn-in step, we simulated the split of African and non-African populations at 55 kya. Following the split, the non-African population experienced a bottleneck with $N_e = 1861$ (as inferred by Gravel et al. (11)). All simulated individuals and accumulated mutations were saved to a population output file for use in the second step.

In the second step, we simulated a single pulse of admixture from Neandertals into the non-African population at a rate of 10%. To track Neandertal ancestry along simulated genomes through time, we placed 50,000 neutral Neandertal markers outside of any potentially functional sequence (which was determined as a union of all annotated Ensembl regions) (Fig. 3A). The locations of these markers were randomly sampled from the set of nearly-fixed Yoruba-Neandertal differences present on the archaic admixture array (SNP panel 4 in (2)). Furthermore, to be able to track Neandertal ancestry within regions directly under negative selection, we placed additional set of fixed Neandertal markers within those regions (Fig. 3A).

Because the efficacy of selection is related to the N_e of the population under consideration (12), we evaluated different demographic models for non-Africans, including a widely-used model by Gravel et al. (11) (i.e. long bottleneck followed by a period of exponential growth), a model of initial slow linear growth post admixture, as well as a model of constant N_e after Neandertal introgression (Fig. S14). However, because we found that the N_e of the admixed non-African population did not have an impact on the slope of the trajectory of Neandertal ancestry over time (Fig. S10), the main results in our paper were performed using a demographic model with constant $N_e = 10,000$.

To track dynamics of selection over time, we periodically saved the simulation state, saving all mutations still segregating at each time-point (both neutral markers and deleterious modern human and Neandertal mutations) in a sample of 500 diploid individuals in VCF format for further analysis. For efficiency reasons, we saved only simulation states in generations 1-10, 20, 50, 100 and then every 200th generation until the final generation 2200 (i.e. 55 thousand years post-introgression, assuming generation time of 25 years).

S4. Evaluating the effect of negative selection against introgression

All of the following analyses were performed on VCF outputs from 20 replicates of our *SLiM* simulations, described in the previous section. Trajectories of Neandertal ancestry in a population over time (Figs. 3B and S9-13) were calculated by averaging the frequencies of all neutral Neandertal markers in a simulation in each time point across 500 sampled diploid individuals. Analysis of the efficacy of selection against introgression as a function of distance from regions carrying deleterious variants (Fig. 3C) was performed by binning the 50,000 neutral Neandertal markers into 5 quintiles, based on their distance from the nearest region under selection. The lowest bin “0” contains neutral Neandertal markers within regions that carried accumulated deleterious mutations. Neandertal ancestry proportions were then calculated for each of the 1,000 sampled chromosomes in each bin, combined from all 20 simulation replicates. Analysis of allele frequency changes over time was performed by calculating the frequency change of each class of variant (neutral Neandertal markers within and outside of selected

regions, and deleterious mutations) between each consecutive pair of sampled time-points, and then averaged over all mutations. For example, if x and y are allele frequencies of a mutation at time-points a and b , then the allele frequency change was calculated as $(x - y) / (a - b)$. This calculation was repeated for all 20 simulation replicates and mean frequency changes were plotted for each replicate separately (Fig. 3D).

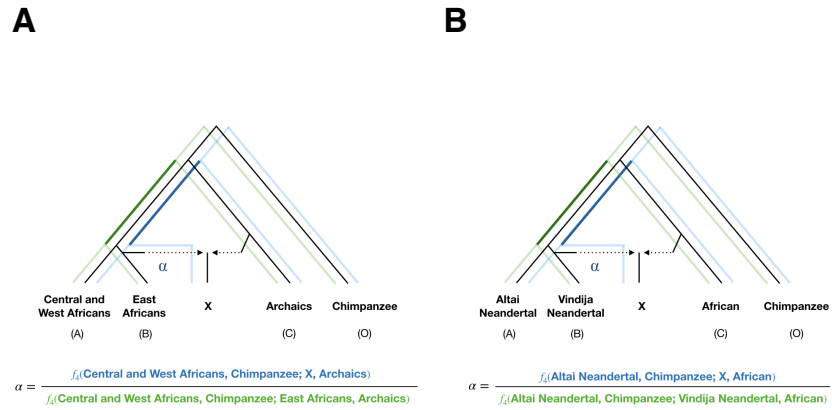


Fig. S1. Tree models underlying indirect and direct f_4 -ratio Neandertal ancestry estimates.

A: Tree model used for the indirect f_4 -ratio. **B:** Tree model used for the direct f_4 -ratio, utilizing two high coverage Neandertal genomes. Blue and green lines represent overlaps of drift paths of f_4 statistics in the numerator and the denominator of f_4 -ratios (13, 14).

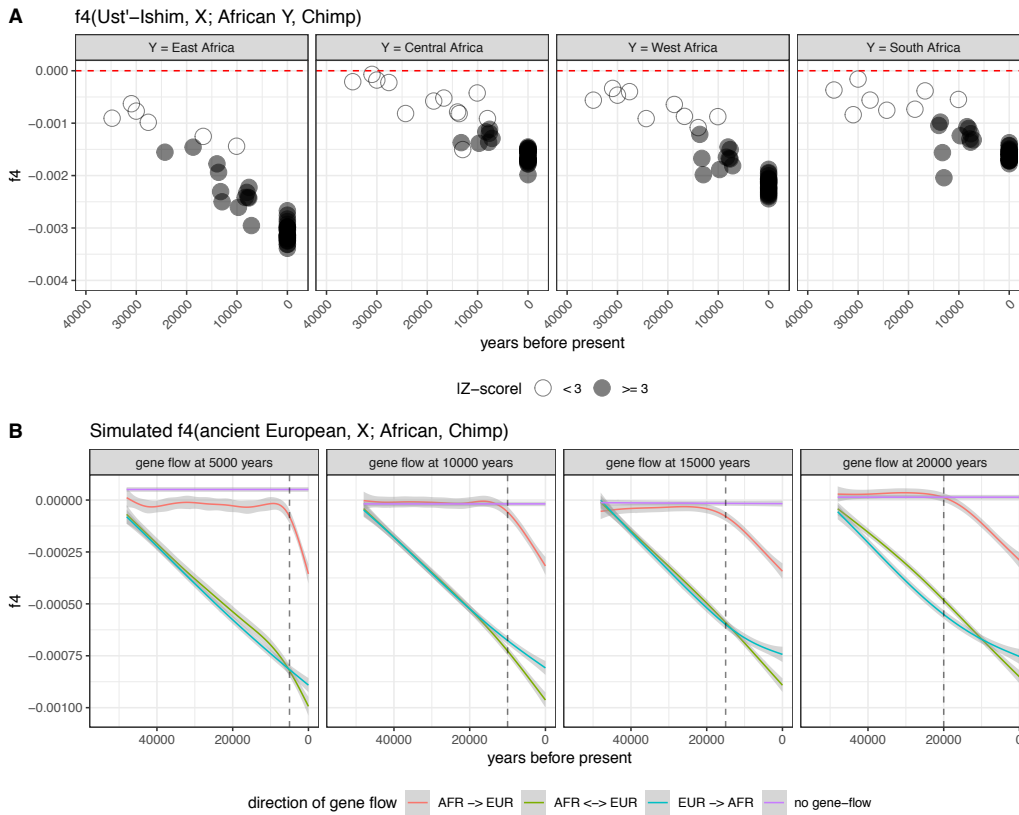


Fig. S2. Increasing affinity between West Eurasians and Africans over time.

A: The statistic $f_4(\text{Ust}'\text{-Ishim, West Eurasian; African Y, Chimp})$ is increasingly negative over time for ancient and present-day West Eurasians (WE), indicating increasing allele sharing (affinity) between WE and Africans with respect to Ust'-Ishim. A variety of demographic forces could cause such shifts, including migration between West Eurasia and Africa, or migration of a third population into both West Eurasia and Africa. The f_4 statistic is not expected to be different from 0 in the absence of admixture. East Africans: Dinka, Bantu, Luhya, Luo, Masai, Somali, West Africans: Esan, Gambian, Mandenka, Mende, Yoruba, Central Africans: Mbuti, Biaka, South Africans: Khomani San, Jul'hoan.

B: Simulations of migration between West Eurasians and Africans starting from 5000 (left) to 20,000 (right) years ago, assuming $N_e = 20000$ in Africans, $N_e = 5000$ in West Eurasians and "total migration" $gm = 0.1$ (g is the duration of gene flow in generations, m is the proportion of the target population composed of migrants in each generation) show

that under a model of migration from West Eurasia to Africa (blue line), this f_4 statistic grows increasingly negative over the past 45ky, regardless of the when the migration event took place, thus making it difficult to determine the timing of such an event.

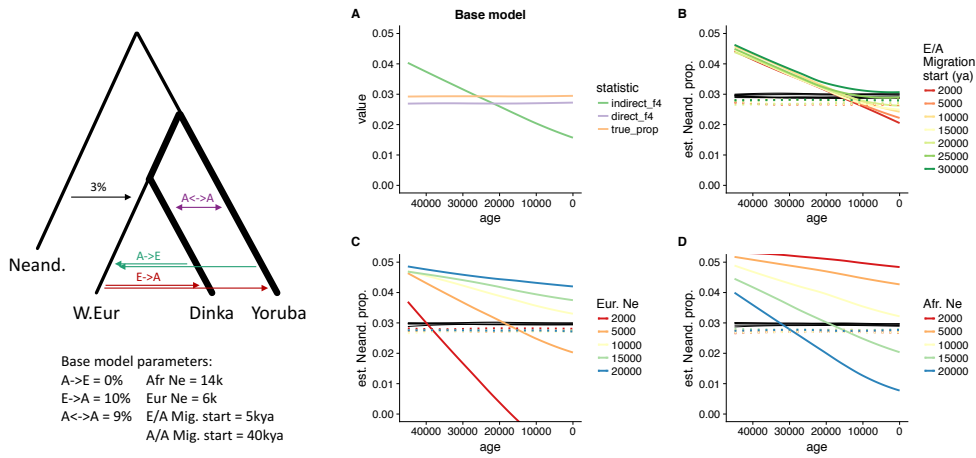


Fig. S3. The effect of N_e and timing on the patterns observed in indirect and direct f_4 -ratio statistics.

A: True Neandertal ancestry, along with direct and indirect f_4 -ratio estimates, on simulated neutral data according to a single demographic model (left) with migration from Europe to Africa, and between two African populations (represented here as Dinka and Yoruba). **B:** Starting from the model in A, the effect of varying the timing of migration on the indirect (solid colored lines), direct (dotted lines), and true Neandertal ancestry proportions (solid black lines). **C:** Similarly, the effect of varying European N_e , and **D)** the effect of varying N_e in both African populations.

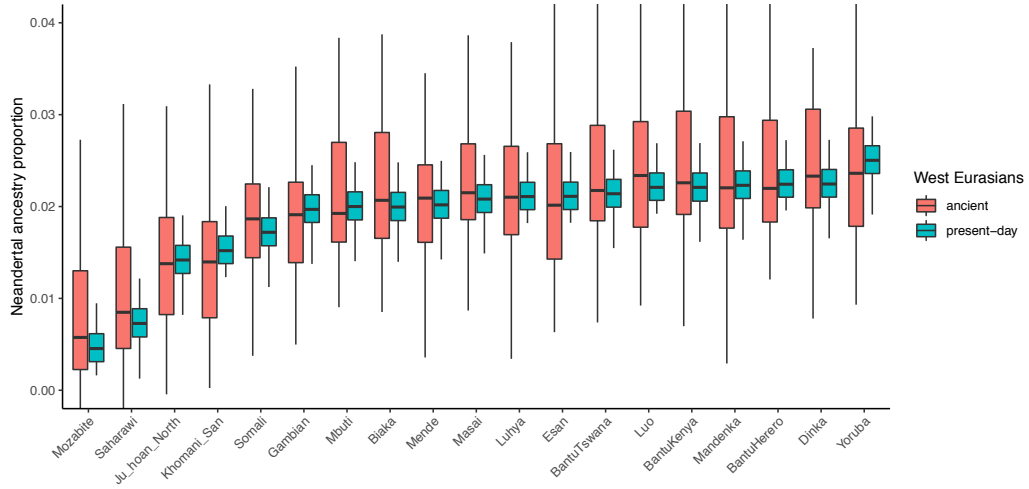


Fig. S4. Neandertal ancestry in ancient and present-day West Eurasians estimated with the direct f_4 -ratio using various African populations.

Neandertal ancestry estimates in ancient and present-day West Eurasians were calculated using the direct f_4 -ratio as $f_4(\text{Altai Neandertal}, \text{Chimp}; X, \text{African}) / f_4(\text{Altai Neandertal}, \text{Chimp}; \text{Vindija Neandertal}, \text{African})$ (Fig. S1). As shown with simulations in Fig. 2, in this statistic, the presence of Neandertal alleles in both Africans and X will cause an underestimate of the true Neandertal ancestry in X , which can be seen in these empirical estimates as well.

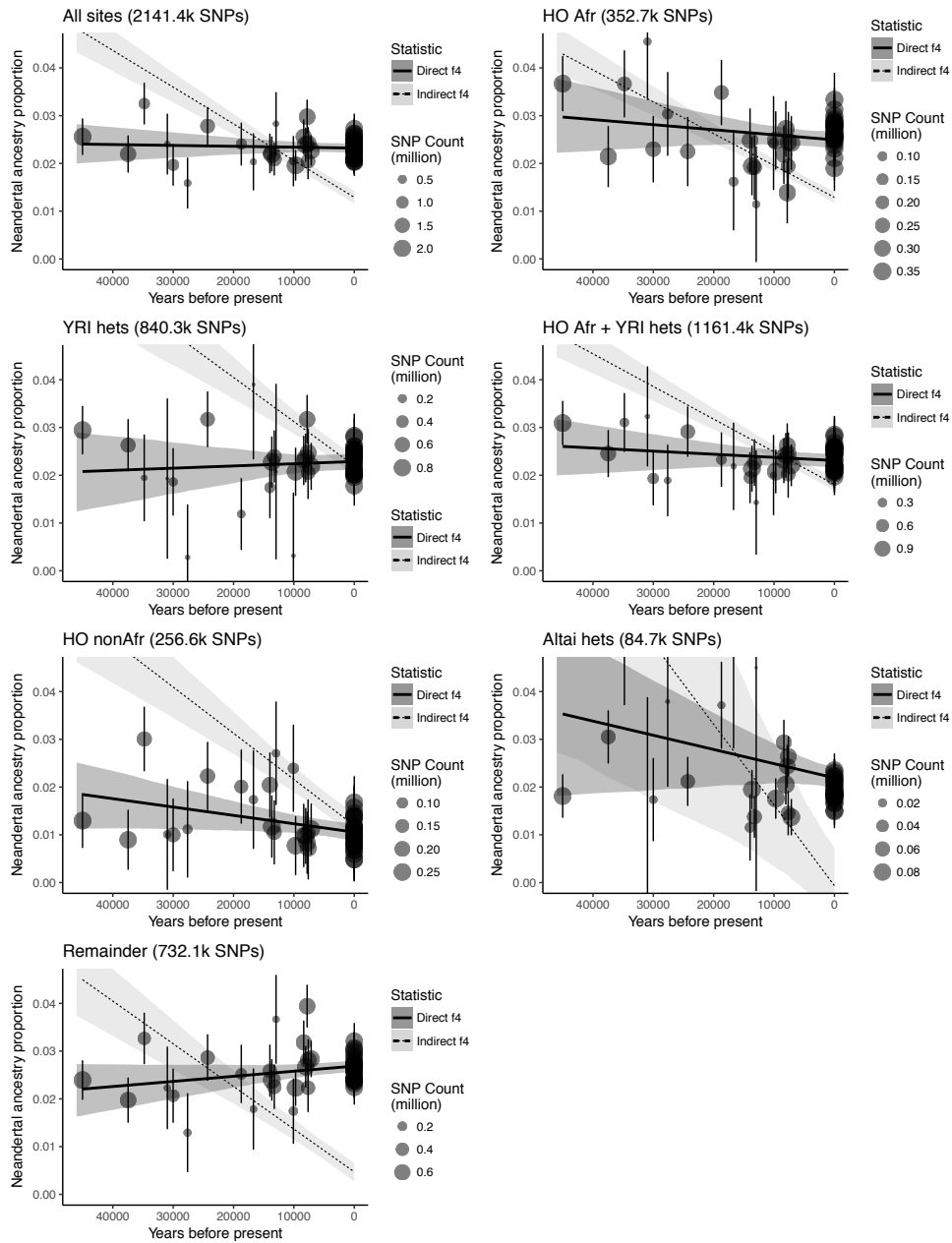


Fig. S5. f_4 -ratio calculations on ascertainment subsets.

Direct and indirect f_4 -ratios are calculated in the same manner as Figure 1, with the data partitioned according to seven ascertainment schemes (from left to right, top to bottom): All 2.2 million SNPs; African ascertained SNPs from the Human Origins array (HO);

Heterozygotes from two Yoruban individuals (YRI hets); Combined African ascertainment (HO nonAfr+ YRI hets); non-African ascertained SNPs from the Human Origins array (HO nonAfr); Heterozygotes in the Altai Neandertal (Altai hets); and the remaining 732k SNPs that do not fit into one of the previous categories. These “remaining” SNPs are largely from the Illumina 610-Quad array and Affymetrix 50k array. Details of these ascertainment can be found in (2).

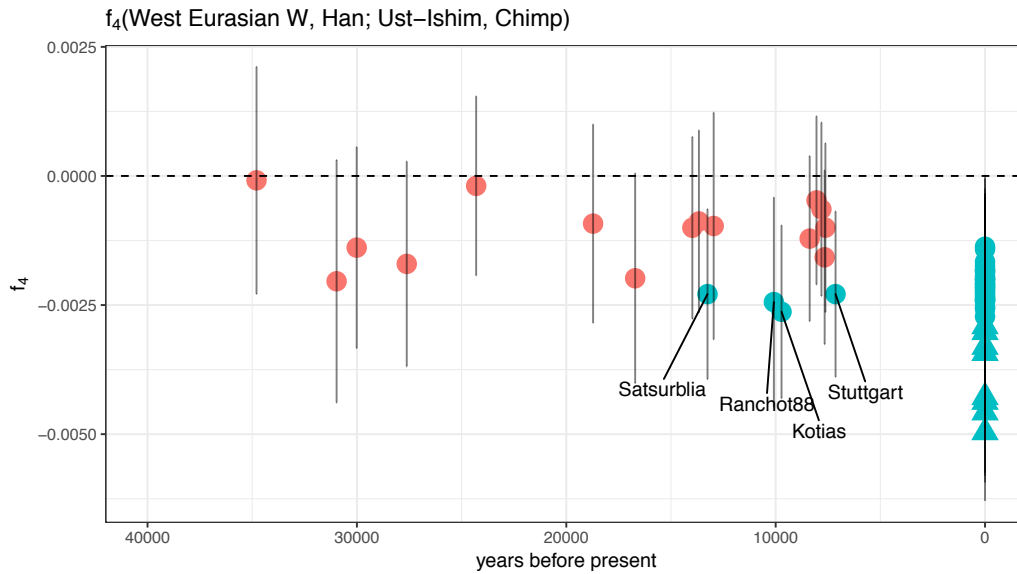


Fig. S6. A signal of Basal Eurasian ancestry in West Eurasia over time.

The statistic $f_4(\text{West Eurasian } W, \text{ Han}; \text{Ust}'\text{-Ishim}, \text{ Chimp})$ has been previously used as a test of the presence of Basal Eurasian ancestry in a West Eurasian W (15). Specifically, it tests whether a population tree in which W and Han lineages form a clade is consistent with the observed data, which results in f_4 statistic ~ 0 . On the other hand, significantly negative values are evidence for an affinity of Han and Ust'-Ishim lineages, which can be most parsimoniously explained by W carrying an ancestry component from a population that diverged from other Eurasians prior to the separation of Ust'-Ishim. This “ghost” population is commonly referred to as Basal Eurasians (16). By analyzing a combined early-modern and present-day West Eurasian dataset, we find that this f_4 statistic becomes consistently negative in the present, which is in agreement with the hypothesis that present-day West Eurasians carry (in different proportions) Basal Eurasian ancestry that was not present in early European hunter gatherers. Blue color indicates individuals with significantly negative f_4 statistic. Present-day individuals are Europeans (circles) and Near Easterners (triangles) from the SGDP panel (17). The SGDP identifiers for Near East individuals used for this grouping are BedouinB, Yemenite_Jew, Palestinian, Iraqi_Jew, Jordanian, Druze, Iranian, Samaritan.

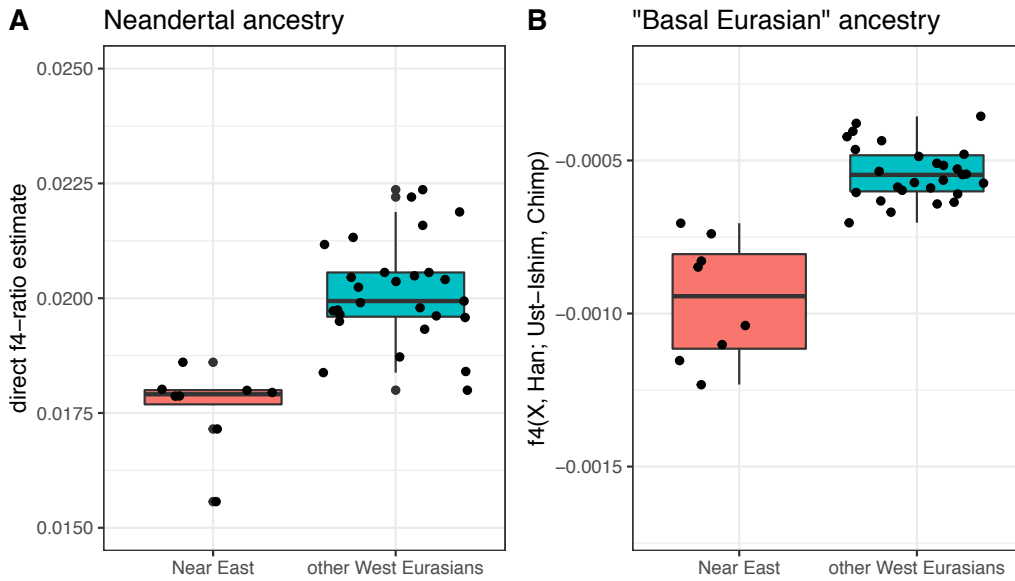


Fig. S7. Proportions of Neandertal ancestry (A) and the amounts of “Basal Eurasian” ancestry (B) in present-day Near Easterners vs other West Eurasians. Panel B shows the same data as the present-day data points in Fig. S6, but is split into two groups – Near Easterners and other West Eurasians. The SGDP identifiers for Near East individuals used for this grouping are BedouinB, Yemenite Jew, Palestinian, Iraqi Jew, Jordanian, Druze, Iranian and Samaritan (17).

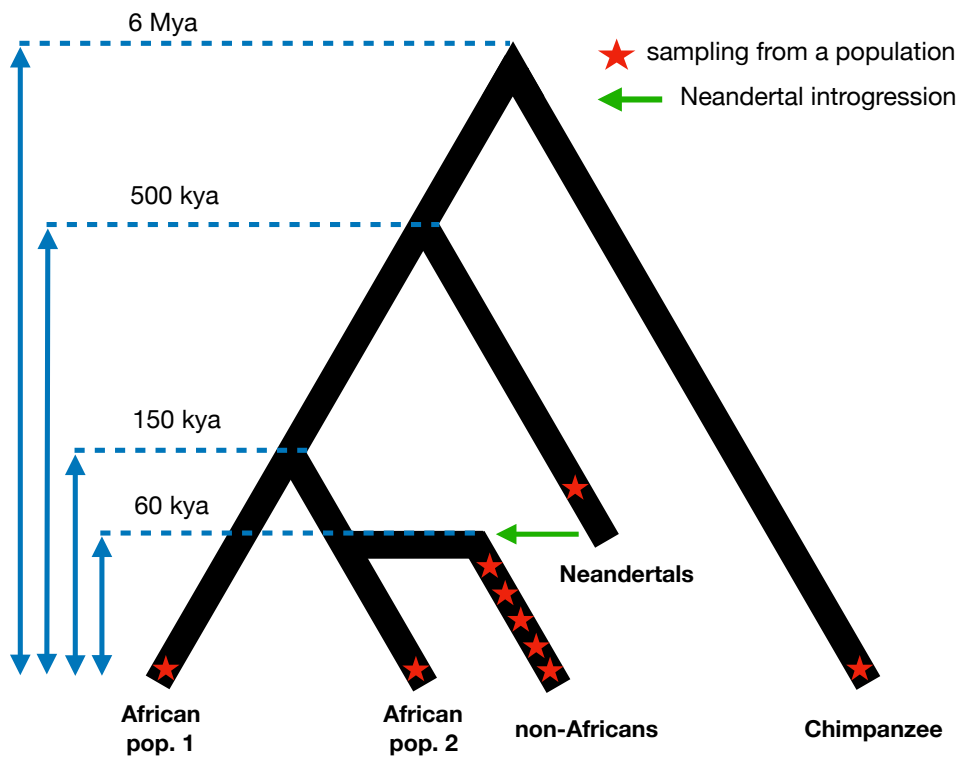


Fig. S8. Demographic model used for testing the temporal behavior of admixture statistics.

Blue dashed lines show split times between simulated populations, red stars indicate approximate points in time at which simulated chromosomes were sampled.

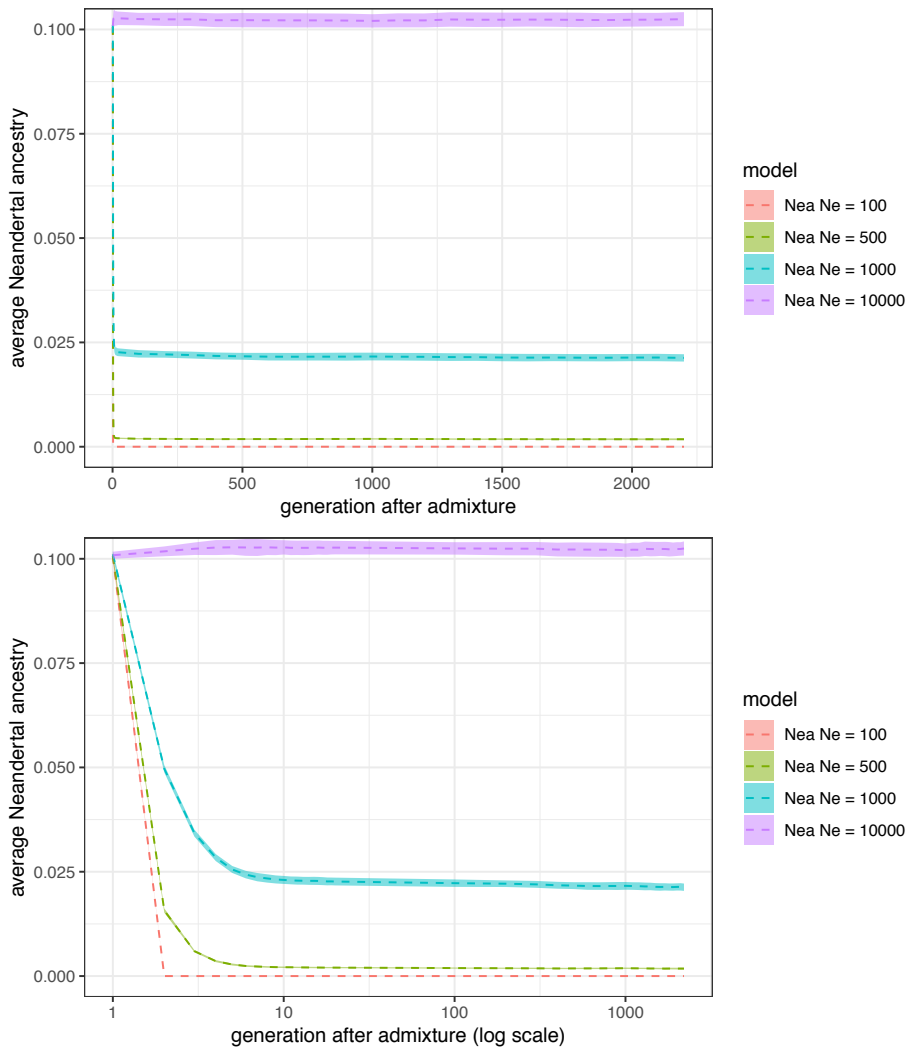


Fig. S9. The effect of Neandertal N_e (Nea N_e) on trajectories of Neandertal ancestry after introgression.

Top and bottom, panels show linear and logarithmic timescales, respectively. The lower the N_e of Neandertal population, the more deleterious alleles behave nearly neutrally, allowing them to reach high frequencies in the Neandertals (4, 18). This imposes a stronger genetic load of the initial modern-human-Neandertal hybrids, causing a more abrupt removal of Neandertal ancestry in the generations shortly after admixture. The shaded areas represent 95% confidence intervals.

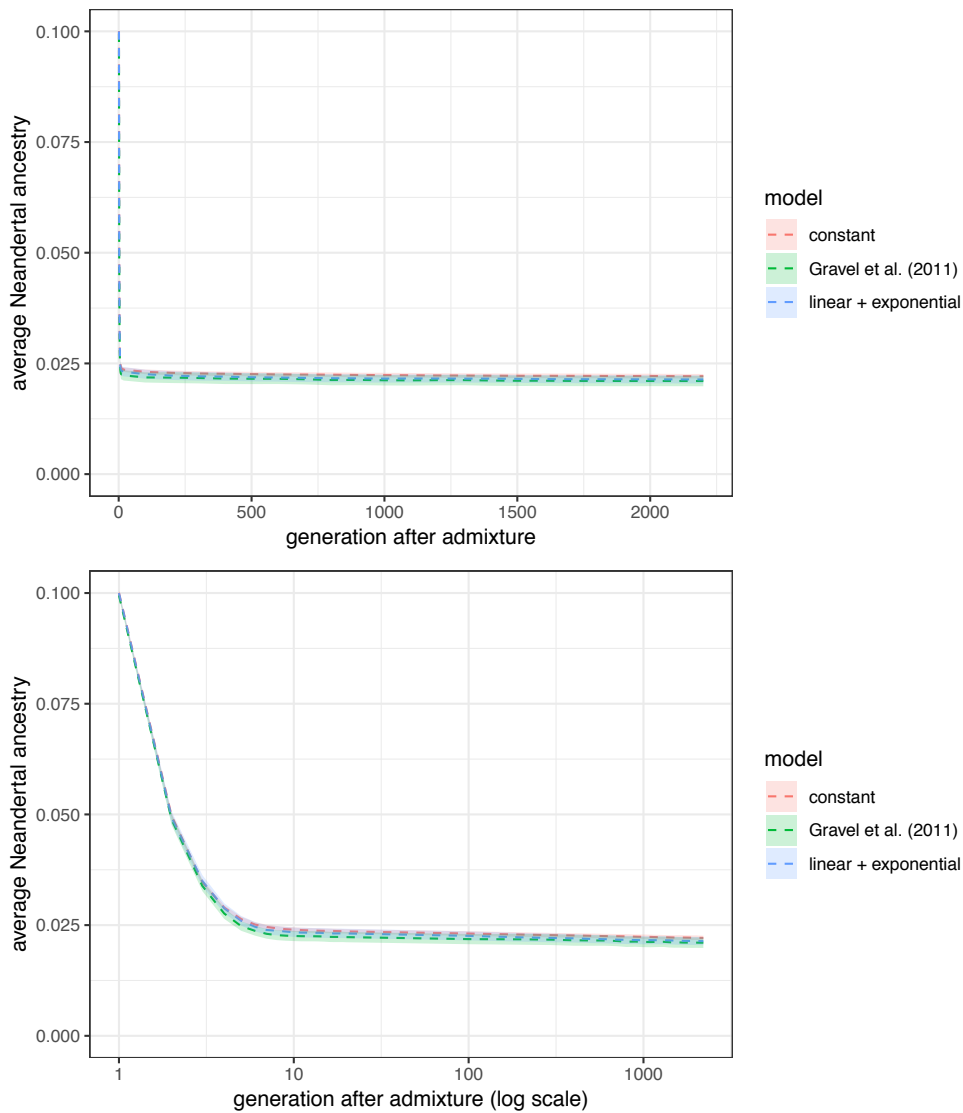


Fig. S10. The effect of non-African demography on trajectories of Neandertal ancestry after introgression.

Top and bottom panels show linear and logarithmic timescales, respectively. Although N_e as a function of time differs dramatically between all three demographic models that we considered (Fig. S14), changing this parameter does not have a strong impact on the overall shape of Neandertal ancestry trajectories. The shaded areas represent 95% confidence intervals.

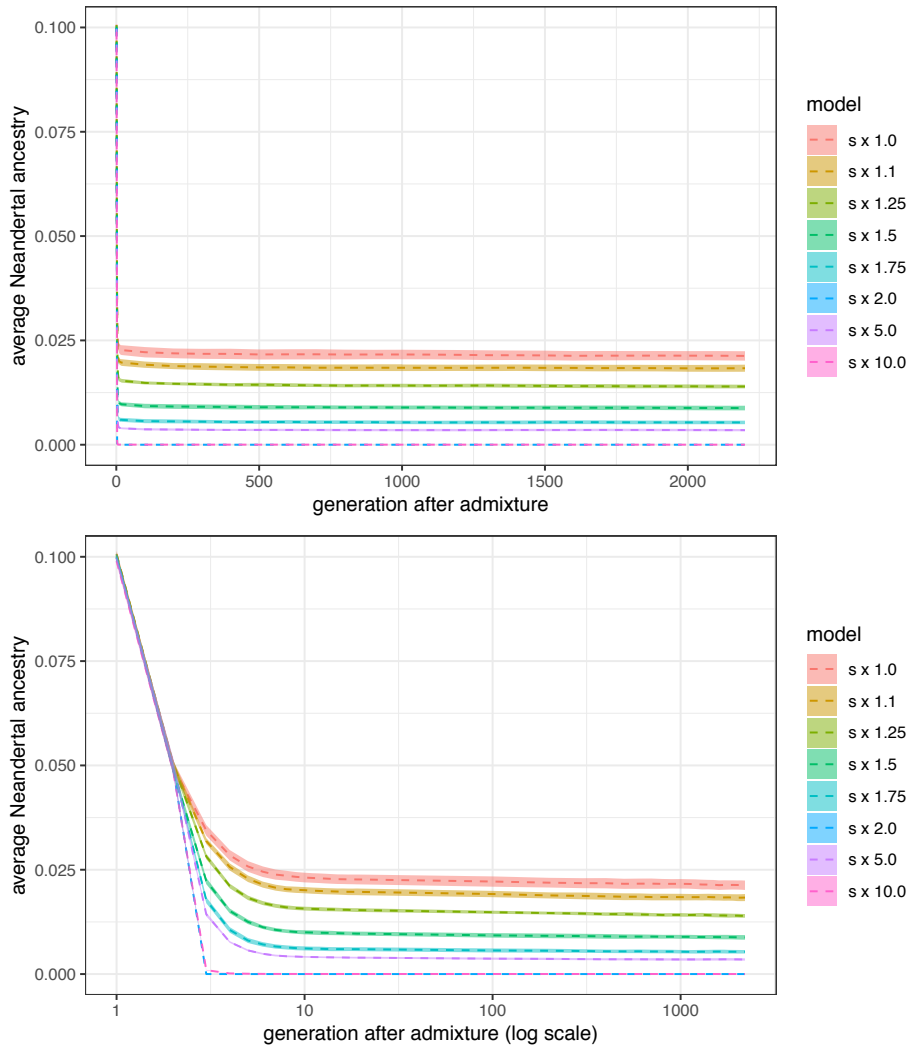


Fig. S11. The effect of making Neandertal mutations more deleterious by increasing their selection coefficients.

Top and bottom panels show linear and logarithmic timescales, respectively. We artificially increased the selection coefficient s of introgressed Neandertal deleterious mutations by multiplying their s by a constant factor. We find that this affects only the final level of Neandertal ancestry in the population, due to stronger genetic burden on hybrids in the first generations after admixture. The shaded areas represent 95% confidence intervals.

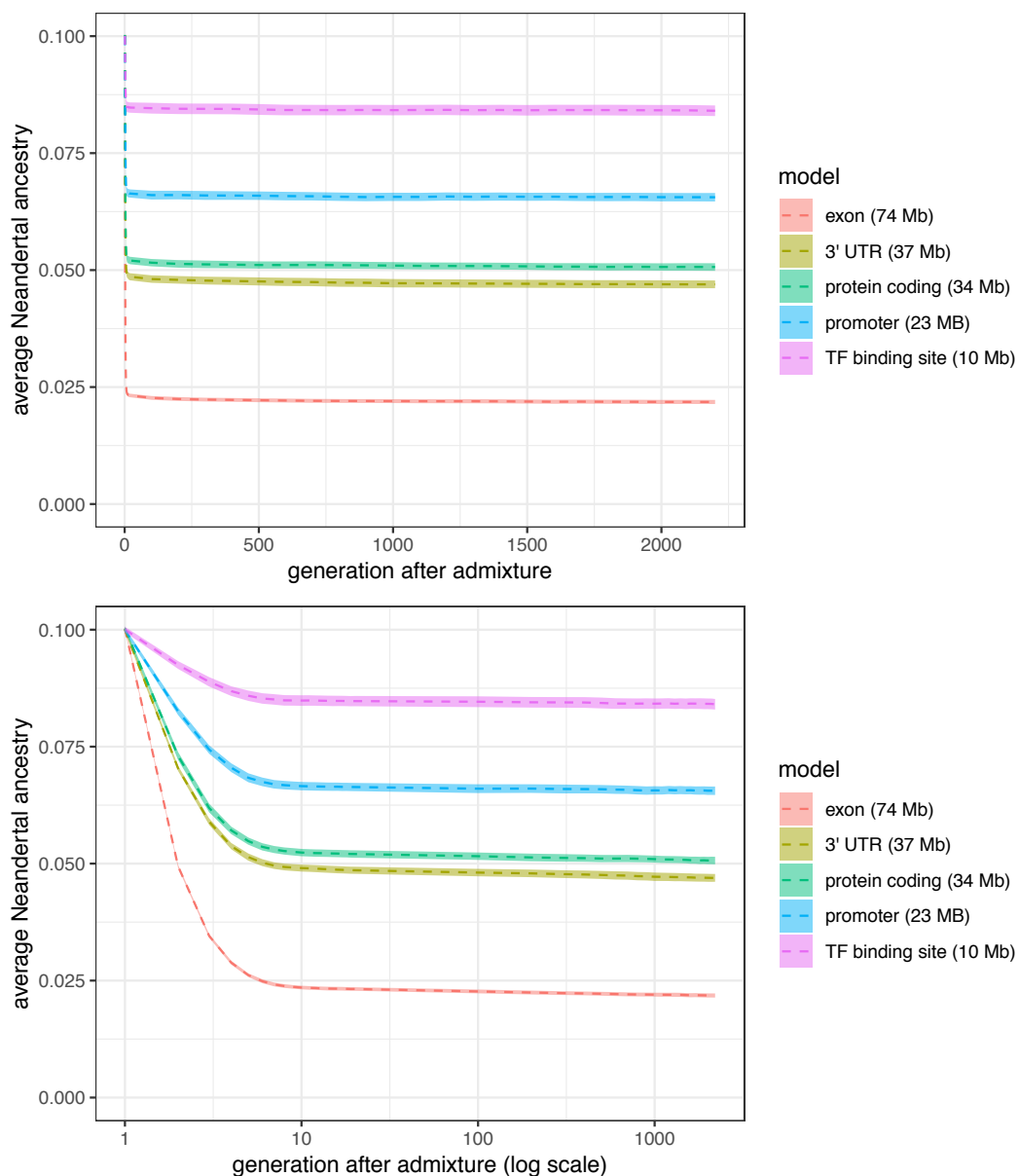


Fig. S12. The effect of changing the total amount of potentially deleterious sequence.

Top and bottom panels show linear and logarithmic timescales, respectively. We simulated deleterious mutations in either full exonic, 3' UTR, protein coding, promoter, or TF binding site regions. Simulations with larger “targets” for deleterious mutations have lower final levels of Neanderthal ancestry. The shaded areas represent 95% confidence intervals.

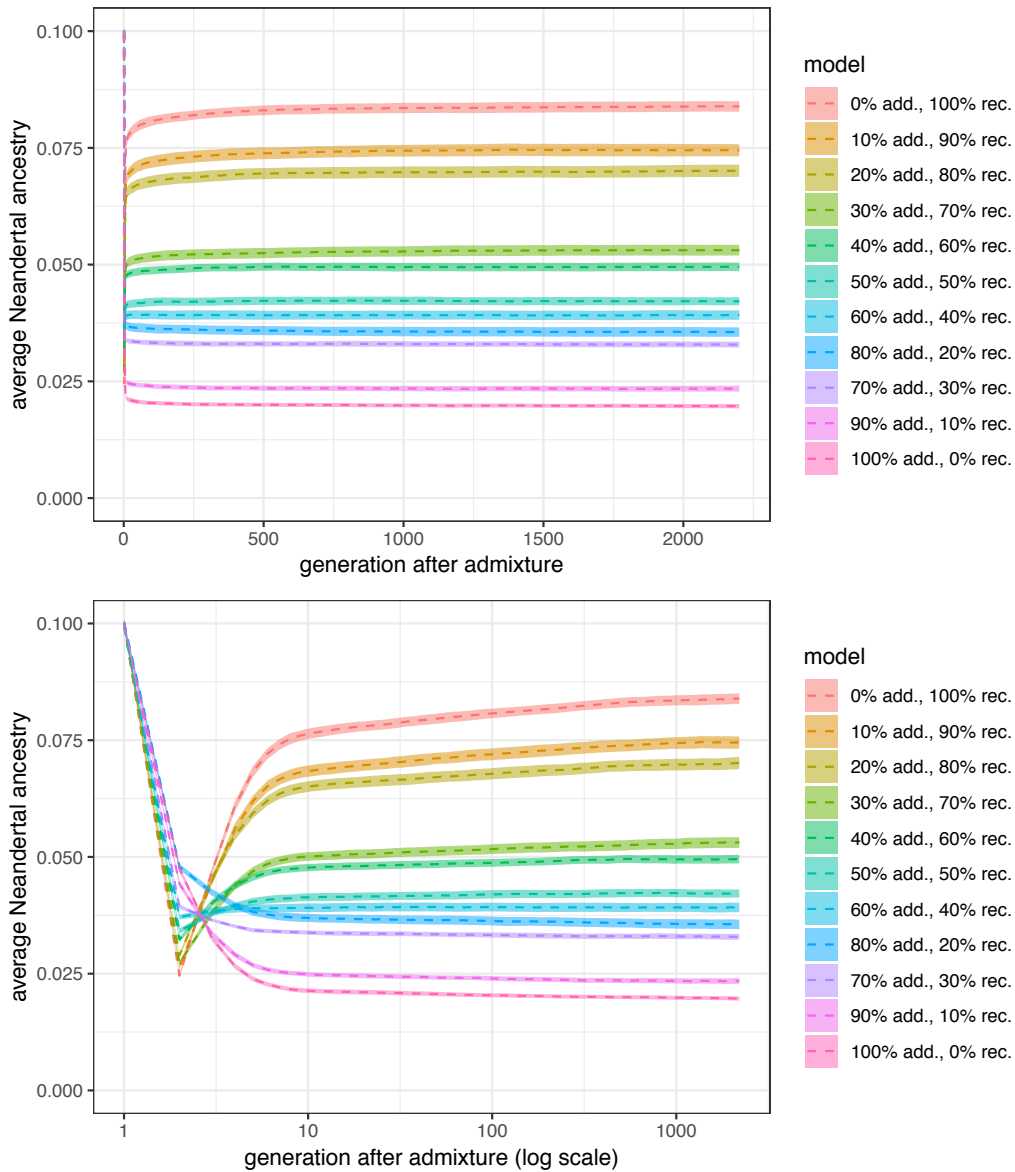


Fig. S13. The effect of changing the proportions of recessive and additive mutations.

Top and bottom panels show linear and logarithmic timescales, respectively. It has been shown that the dominance coefficient of deleterious mutations can lead to Neandertal ancestry trajectories following entirely opposite patterns (4). Specifically, models with only recessive mutations lead to an initial increase of the Neandertal ancestry proportions due to heterosis (4). Due to these opposing effects of dominance, we investigated

scenarios with different mixtures of dominance coefficients of deleterious mutations. We found that changing the ratios of recessive and additive mutations affects only the final baseline of Neandertal ancestry in the population, and does not lead to a steady decline in Neandertal ancestry over time. The shaded areas represent 95% confidence intervals.

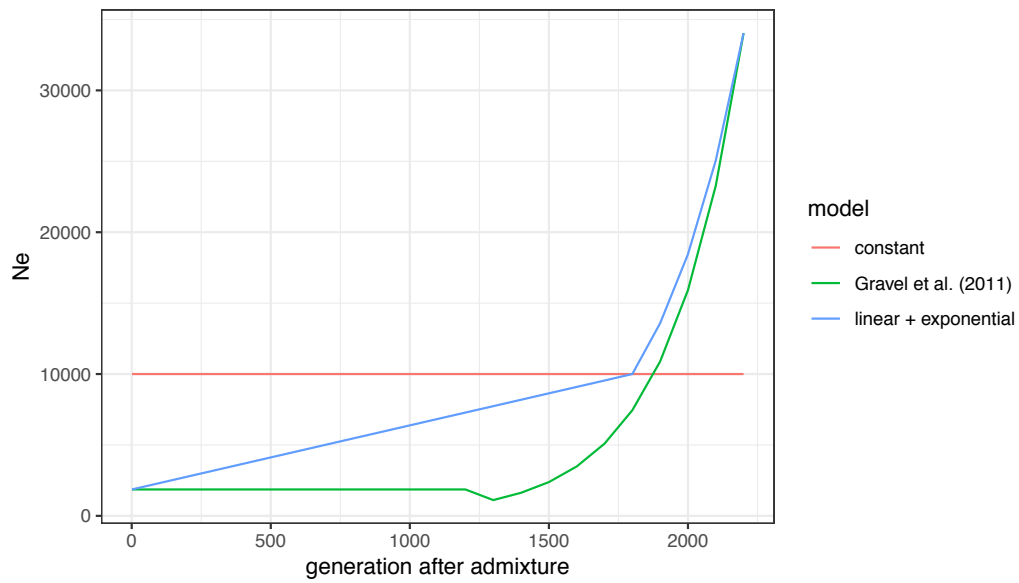


Fig. S14. Three models of non-African demography after Neandertal admixture.

N_e as a function of time for three models of non-African demography: a model of constant N_e after Neandertal introgression, a model of initial slow linear growth post admixture, and a long bottleneck followed by a period of exponential growth (Gravel et al. (11)). Unless otherwise noted, all analyses in this paper used the constant N_e model.

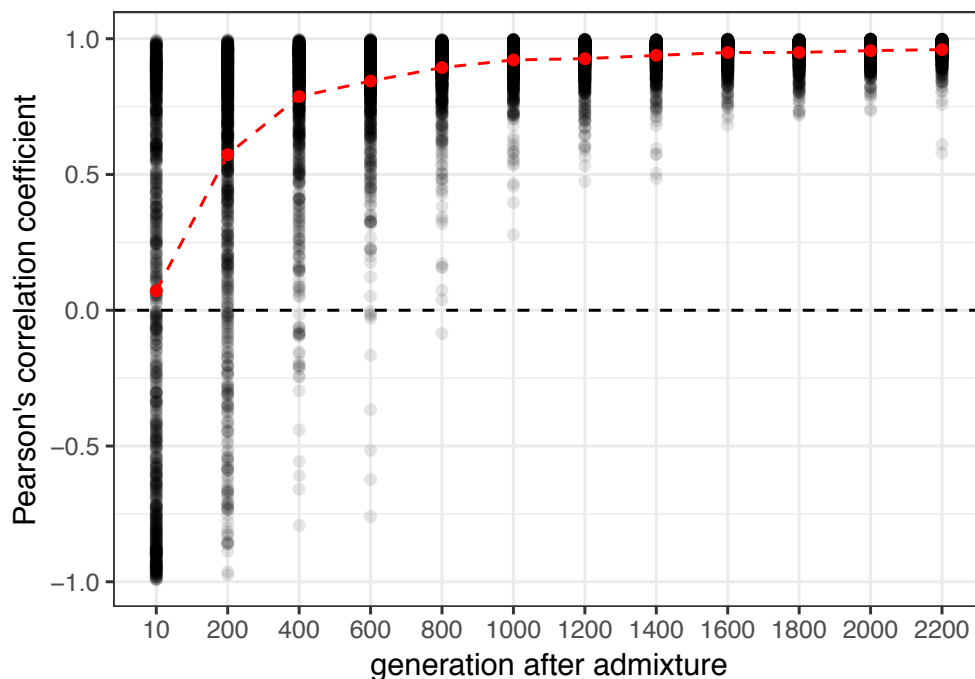


Fig. S15. Coefficients of correlation between the proportion of surviving Neandertal ancestry and distance to a genomic region under negative selection.

Pearson's correlation coefficient of a correlation between Neandertal ancestry proportion and the distance to the nearest region under negative selection, at a given point in time after introgression. Each black dot represents a correlation coefficient in a single simulated "individual" with 500,000 informative sites. Red dots indicate mean correlation coefficient at a given time-point. This figure uses the same data presented in Fig. 3C (which shows averages over all simulations at each individual time-point).

Table S1. Proportions of overlapping functional categories.

Each row contains proportions of overlap of a given region (row label) with all other annotated regions (column labels). CDS: protein coding sequence.

	protein coding	5' UTR	3' UTR	enhancer	promoter	phastCons
protein coding		0.050	0.174	0.003	0.051	0.615
5' UTR	0.163		0.039	0.005	0.402	0.244
3' UTR	0.157	0.011		0.007	0.009	0.256
enhancer	0.005	0.003	0.013		0.000	0.080
promoter	0.076	0.183	0.015	0.000		0.153
phastCons	0.206	0.025	0.095	0.000	0.034	

References

1. Kelleher J, Etheridge AM, McVean G (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* 12(5):e1004842.
2. Fu Q, et al. (2015) An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524(7564):216–219.
3. Haller BC, Messer PW (2017) SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol Biol Evol* 34(1):230–240.
4. Harris K, Nielsen R (2016) The Genetic Cost of Neanderthal Introgression. *Genetics:genetics*.116.186890.
5. Zerbino DR, et al. (2018) Ensembl 2018. *Nucleic Acids Res* 46(D1):D754–D761.
6. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.
7. Haeussler M, et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*. doi:10.1093/nar/gky1095.
8. International HapMap Consortium, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.
9. Eyre-Walker A, Woolfit M, Phelps T (2006) The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics* 173(2):891–900.
10. Vernot B, Akey JM (2014) Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* 343(6174):1017–1021.
11. Gravel S, et al. (2011) Demographic history and rare allele sharing among human populations. *PNAS* 108(29):11983–11988.
12. Lanfear R, Kokko H, Eyre-Walker A (2014) Population size and the rate of evolution. *Trends in Ecology & Evolution* 29(1):33–41.
13. Patterson N, et al. (2012) Ancient Admixture in Human History. *Genetics* 192(3):1065–1093.
14. Peter BM (2016) Admixture, Population Structure and F-Statistics. *Genetics:genetics*.115.183913.
15. Lazaridis I, et al. (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature* 536(7617):419–424.

16. Lazaridis I, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513(7518):409–413.
17. Mallick S, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538(7624):201–206.
18. Juric I, Aeschbacher S, Coop G (2016) The Strength of Selection against Neanderthal Introgression. *PLOS Genetics* 12(11):e1006340.

Chapter 2

The evolutionary history of Neanderthal and Denisovan Y chromosomes

Petr, M., Hajdinjak, M., Fu, Q., Essel, E., Rougier, H., Crevecoeur, I., Semal, P., Golovanova, L. V., Doronichev, V. B., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Shunkov, M. V., Kozlikin, M. B., Derevianko, A. P., Vernot, B.; Meyer, M., Kelso, J.

Science, 369, 1653–1656 (2020)

RESEARCH

Y EVOLUTION

The evolutionary history of Neanderthal and Denisovan Y chromosomes

Martin Petr^{1*}, Mateja Hajdinjak^{1,2}, Qiaomei Fu^{3,4,5}, Elena Essel¹, H el ene Rougier⁶, Isabelle Crevecoeur⁷, Patrick Sema⁸, Liubov V. Golovanova⁹, Vladimir B. Doronichev⁹, Carles Lalueza-Fox¹⁰, Marco de la Rasilla¹¹, Antonio Rosas¹², Michael V. Shunkov¹³, Maxim B. Kozlikin¹³, Anatoli P. Derevianko¹³, Benjamin Vernot¹, Matthias Meyer¹, Janet Kelso^{1*}

Ancient DNA has provided new insights into many aspects of human history. However, we lack comprehensive studies of the Y chromosomes of Denisovans and Neanderthals because the majority of specimens that have been sequenced to sufficient coverage are female. Sequencing Y chromosomes from two Denisovans and three Neanderthals shows that the Y chromosomes of Denisovans split around 700 thousand years ago from a lineage shared by Neanderthals and modern human Y chromosomes, which diverged from each other around 370 thousand years ago. The phylogenetic relationships of archaic and modern human Y chromosomes differ from the population relationships inferred from the autosomal genomes and mirror mitochondrial DNA phylogenies, indicating replacement of both the mitochondrial and Y chromosomal gene pools in late Neanderthals. This replacement is plausible if the low effective population size of Neanderthals resulted in an increased genetic load in Neanderthals relative to modern humans.

Ancient DNA (aDNA) has transformed our understanding of human evolutionary history, revealing complex patterns of population migration and gene flow, including admixture from archaic humans into modern humans. Particularly important have been analyses of autosomal sequences (1, 2), which represent a composite of genealogies of any individual's ancestors. Although mitochondrial DNA (mtDNA) and Y chromosomes only provide information about single maternal and paternal lineages, they offer a distinctive perspective on various aspects of population history such as sex-specific migration, matrilocality and patrilocality, and variance in reproductive success between individuals (3–5). Furthermore, because of their lower effective population size (N_e) compared with that of autosomal loci, coalescent times of mtDNA and Y chromosomes sampled from two populations provide an upper bound for the last time they experienced gene flow.

¹Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. ²The Francis Crick Institute, NW1 1AT London, UK. ³Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing 100044, China. ⁴CAS Center for Excellence in Life and Paleoenvironment, Beijing 100044, China. ⁵University of Chinese Academy of Sciences, Beijing 100049, China. ⁶Department of Anthropology, California State University, Northridge, Northridge, CA 91330-8244, USA. ⁷Universit e de Bordeaux, CNRS, UMR 5199-PACEA, 33615 Pessac Cedex, France. ⁸Royal Belgian Institute of Natural Sciences, 1000 Brussels, Belgium. ⁹ANO Laboratory of Prehistory 14 Linia 3-11, St. Petersburg 1990 34, Russia. ¹⁰Institute of Evolutionary Biology, Consejo Superior de Investigaciones Cientificas, Universitat Pompeu Fabra, 08003 Barcelona, Spain. ¹¹Area de Prehistoria, Departamento de Historia, Universidad de Oviedo, 33011 Oviedo, Spain. ¹²Departamento de Paleobiologia, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Cientificas, 28006 Madrid, Spain. ¹³Institute of Archaeology and Ethnography, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia. *Corresponding author. Email: mp@bodk.net (M.P.); kelso@eva.mpg.de (J.K.)

The mtDNA and autosomal sequences of Neanderthals, Denisovans, and modern humans have revealed puzzling phylogenetic discrepancies. Autosomal genomes show that Neanderthals and Denisovans are sister groups that split from modern humans between 550 thousand and 765 thousand years (ka) ago (6). By contrast, the mtDNAs of Neanderthals and modern humans are more similar to one another [time to the most recent common ancestor (TMRCA) of 360 to 468 ka ago] than to the mtDNAs of Denisovans (7). Notably, ~400-ka-old early Neanderthals from Sima de los Huesos were shown to carry mitochondrial genomes related to Denisovan mtDNAs (8, 9). This suggests that Neanderthals originally carried a Denisovan-like mtDNA, which was later completely replaced through ancient gene flow from an early lineage related to modern humans (7, 9).

The Y chromosomes of Neanderthals and Denisovans should provide an additional source of information about population splits and gene flow events between archaic and modern humans or populations related to them. However, with the exception of a small amount of Neanderthal Y chromosome coding sequence (118 kb) (10), none of the male Neanderthals or Denisovans studied to date have yielded sufficient amounts of endogenous DNA to allow comprehensive studies of archaic human Y chromosomes.

Previous genetic studies identified two male Denisovans, Denisova 4 (55 to 84 ka old) and Denisova 8 (106 to 136 ka old) (11, 12), and two male late Neanderthals, Spy 94a (38 to 39 ka old) and Mezmaiskaya 2 (43 to 45 ka old) (13) (Fig. 1A). To enrich for Y chromosome DNA from these individuals, we performed hybridization capture using probes we designed to target ~6.9 Mb of the nonrecombining portion

of the human Y chromosome (Fig. 1B) (14). This yielded sequence coverage of 1.4× for Denisova 4, 3.5× for Denisova 8, 0.8× for Spy 94a, and 14.3× for Mezmaiskaya 2 (Fig. 1C and table S2). In addition, we used a capture array designed for modern human Y chromosomes (3) to obtain 7.9× coverage of ~560 kb of the Y chromosome from the ~46- to 53-ka-old El Sidr on 1253 Neanderthal (Fig. 1C and table S2), which has been analyzed previously (15, 16).

To call genotypes of the captured archaic human and previously published modern human Y chromosomes (4, 17, 18), we leveraged the haploid nature of the human Y chromosome and implemented a consensus approach that requires at least 90% of the reads observed at each site covered by at least three reads to agree on a single allele (14). This minimizes the impact of aDNA damage on genotyping accuracy while allowing for a small amount of sequencing error or contamination (fig. S8) (14).

To determine the relationships between Denisovan, Neanderthal, and modern human Y chromosomes, we constructed a neighbor-joining tree from the Y chromosome genotype calls (14). Unlike the rest of the nuclear genome, which puts Denisovans and Neanderthals as sister groups to modern humans (2), the Denisovan Y chromosomes form a separate lineage that split before Neanderthal and modern human Y chromosomes diverged from each other (Fig. 2A). Notably, all three late Neanderthal Y chromosomes cluster together and fall outside of the variation of present-day human Y chromosomes (Fig. 2A).

To estimate the TMRCA of archaic and modern human Y chromosomes, we adapted a previously published method that calculates the archaic-modern human TMRCA as a proportion of the deepest known split in present-day human Y variation (4, 10, 14) and is therefore robust to low coverage and aDNA damage (10) (fig. S8 and table S2). We first calculated the mutation rate in the 6.9-Mb target region to be 7.34×10^{-10} per base pair per year [bootstrap confidence interval (CI) 6.27×10^{-10} to 8.46×10^{-10}] (fig. S11 and table S11) (14) and used it to estimate a TMRCA of ~249 ka ago (bootstrap CI 213 to 293 ka ago) (fig. S11 and table S11) (14) for the African A00 lineage and a set of non-African Y chromosomes (4, 18). This is consistent with other studies of present-day human Y chromosomes (4, 17), suggesting that the Y chromosomal regions we sequenced are not unusual in terms of their mutation rate. We then used this A00 divergence time of 249 ka ago to infer TMRCA between archaic Y chromosomes and present-day non-African Y chromosomes for each archaic individual (fig. S14 and table S12) (14). The two Denisovan Y chromosomes split from the modern human lineage around 700 ka ago (Denisova 8: 707 ka ago, CI 607 to 835 ka ago; Denisova 4: 708 ka ago, CI 550 to 932 ka ago) (Fig. 2B and table

RESEARCH | REPORT

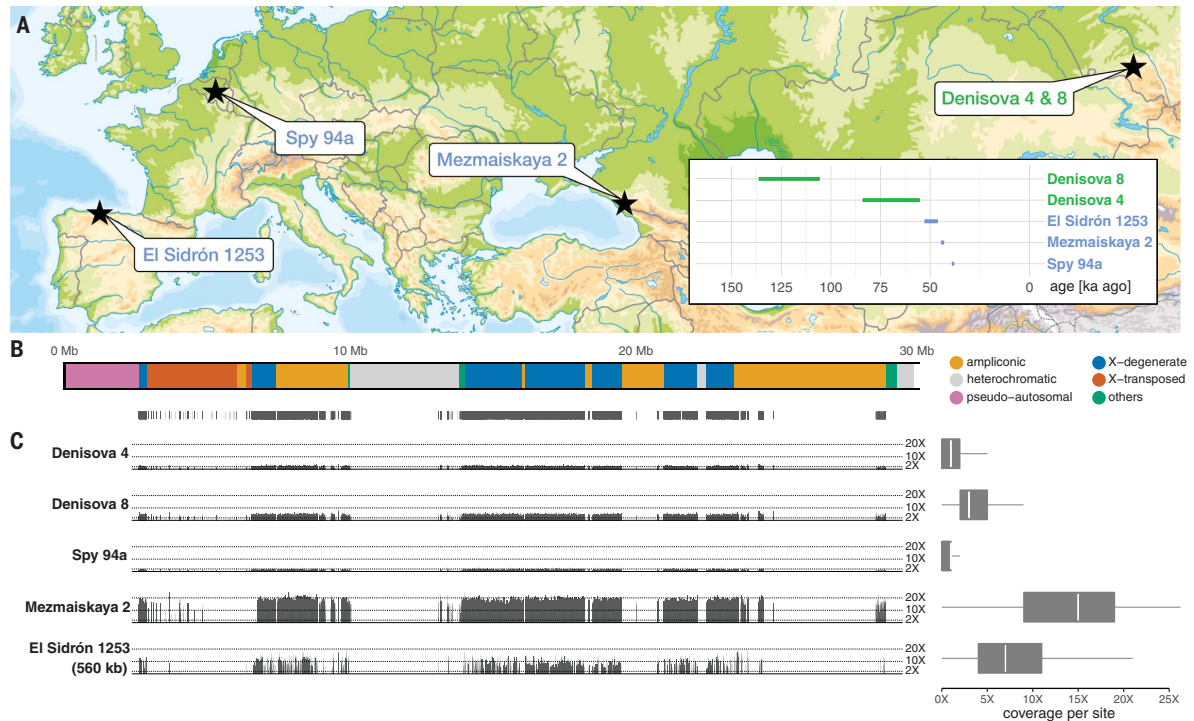


Fig. 1. Overview of male archaic humans in our study. (A) Archaeological site locations. Ages of specimens are shown as an inset (12, 13, 15). (B) Portion of the human Y chromosome targeted for capture [legend on right, coordinates of genomic regions are from (30)]. Thin black vertical lines show individual target capture regions.

(C) (Left) Spatial distribution of sequencing coverage along the ~6.9 Mb of capture target regions. The heights of the thin vertical bars represent average coverage in each target region. Coordinates are aligned to match the chromosome shown in (B). (Right) Coverage across all target sites for each individual to the left.

S12). By contrast, the three Neanderthal Y chromosomes split from the modern human lineage about 370 ka ago: 353 ka ago for Spy 94a (CI 287 to 450 ka ago), 370 ka ago for Mezmaiskaya 2 (CI 326 to 420 ka ago), and 339 ka ago for El Sidrón 1253 (CI 275 to 408 ka ago) (Fig. 2B and table S12). Additionally, we used the proportion of sharing of derived alleles with the high-coverage Mezmaiskaya 2 to estimate the TMRCA of the three Neanderthal Y chromosomes to around 100 ka ago (figs. S25 and S26). We validated the robustness of all TMRCA estimates using filters of varying levels of stringency and different genotype calling methods and also by comparing capture and shotgun sequence results (figs. S19, S21, and S23). Although there was some evidence of capture bias in the data (fig. S7), we observed no consistent differences between capture data and shotgun sequences or between individuals showing different read length distributions, indicating that technical biases do not affect our inferences (fig. S21).

Our estimates of the Neanderthal–modern human TMRCA (Fig. 2B) are younger than the previous estimate of ~588 ka ago from the El Sidrón 1253 individual (10). This older es-

timate was calculated from ~3 \times coverage of 118 kb of nuclear exome capture sequence and, because of the limited amount of data, used single-nucleotide polymorphisms supported even by single reads (10, 16). However, this is problematic because it can result in an increased rate of erroneously called genotypes, leading to some derived alleles that El Sidrón 1253 shares with both Neanderthals and modern humans being converted to the ancestral state, increasing the apparent TMRCA. When we applied filtering designed to mitigate errors (14) to the original El Sidrón 1253 data, we arrived at TMRCA estimates for El Sidrón 1253 consistent with all other Neanderthals in our study (fig. S22).

The Denisovan–modern human Y chromosome TMRCA estimates agree with population split times inferred from autosomal sequences, suggesting that the differentiation of Denisovan Y chromosomes from modern humans occurred through a simple population split (19). By contrast, the young TMRCA of Neanderthal and modern human Y chromosomes and mtDNAs suggest that these loci have been replaced in Neanderthals through gene flow from an early lineage closely related to modern humans

(Fig. 3A) (7). Previous work indicates that the rate of gene flow from modern humans into Neanderthals was on the order of only a few percent (20, 21). Because the fixation probability of a locus is equal to its initial frequency in a population (22), the joint probability of both Neanderthal mtDNA and Y chromosomes being replaced by their introgressed modern human counterparts starting from a low initial frequency is even lower. However, owing to their low N_e and reduced efficacy of purifying selection, Neanderthals have been shown to have accumulated an excess of deleterious variation compared with modern humans (16), and it has been suggested that introgressed DNA was not neutral (23, 24).

To explore the dynamics of modern human Y chromosomes introgressed into Neanderthals, we simulated introgression of a nonrecombining, uniparental locus under purifying selection (14, 25). We considered a range of values for the following parameters: Neanderthal and modern human N_e , the time that both populations evolved independently after their split, and the amount of sequence under selection, all of which affect the amount of deleterious variation that accumulated in Neanderthal

RESEARCH | REPORT

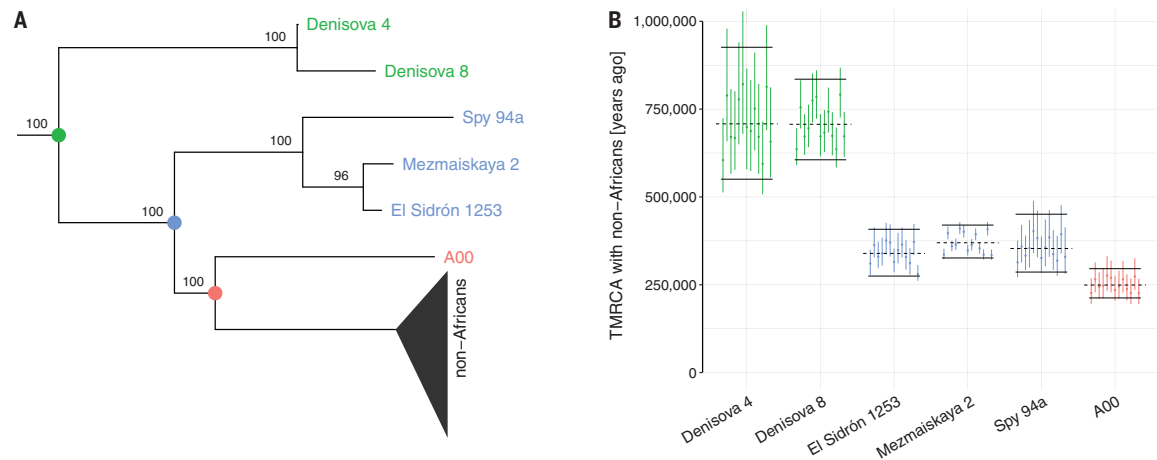


Fig. 2. Phylogenetic relationships between archaic and modern human Y chromosomes. (A) Neighbor-joining tree estimated from the Y chromosome genotype calls, excluding C-to-T and G-to-A polymorphisms, rooted with a chimpanzee as the outgroup (14). Numbers show bootstrap support out of 100 bootstrap replicates. Terminal branch lengths are not informative about the ages of specimens (Fig. 1A), owing to differences in sequence quality.

(B) Estimates of TMRCA between Y chromosomes along the x axis and a panel of 13 non-African Y chromosomes. Each dot represents the TMRCA with a single non-African Y chromosome, with error bars showing 95% CI from a resampling of branch counts (14). Black horizontal lines show the mean TMRCA calculated across the full non-African panel (dashed lines) with resampling-based 95% CI (solid lines) (14).

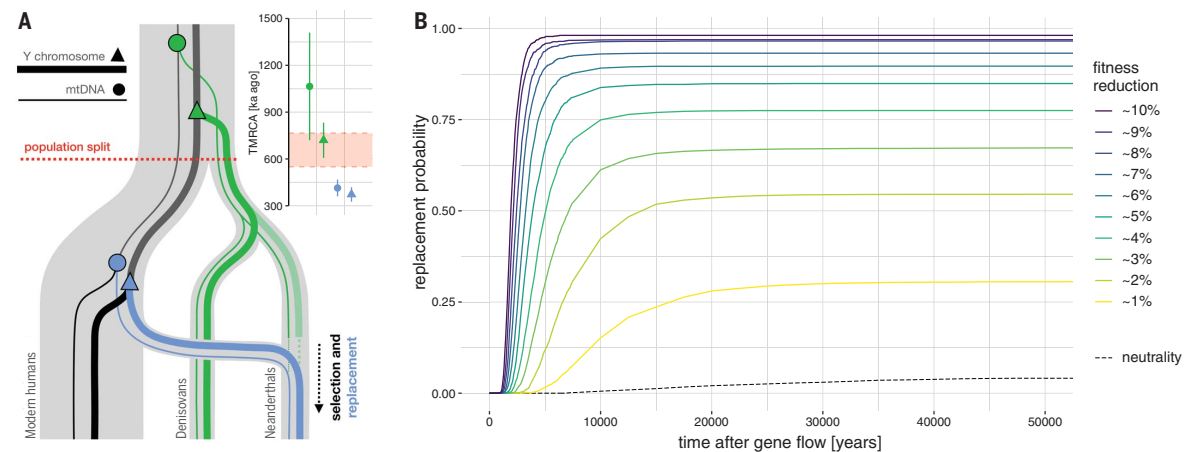


Fig. 3. Proposed model for the replacement of Neanderthal Y chromosomes and mtDNA. (A) Relationships between archaic and modern human mtDNA and Y chromosomes. The semitransparent Neanderthal lineage indicates a (as yet unsampled) hypothetical Y chromosome replaced by an early lineage related to modern human Y chromosomes. Most recent common ancestors with modern human lineages are shown for mtDNA (circles) and Y chromosomes (triangles). The inset shows TMRCA estimates for the four nodes in the diagram: Y chromosome TMRCA as estimated by our study and mtDNA TMRCA estimates from the literature (7, 8). The red shaded area highlights the 95% CI for the

population split time between archaic and modern humans, shown as the dotted red horizontal line (6). (B) Probability of replacement of a non-recombining, uniparental Neanderthal locus over time, assuming a given level of fitness burden relative to its modern human counterpart. Trajectories are based on forward simulations across a grid of parameters (figs. S27 to S29) (14), with N_e of modern humans and Neanderthals fixed at 10,000 and 1000, respectively. Modern human introgression was simulated in a single pulse at 5%. Replacement probabilities from a wider range of model parameters are shown in fig. S31.

and modern human populations before introgression (14). We simulated introgression of modern human Y chromosomes into the Neanderthal population in a single pulse and varied the contribution between 1 and 10%. We then traced the frequency of the intro-

gressed modern human Y chromosomes in Neanderthals over 100 ka. For each combination of parameters, we calculated how much lower the fitness of an average Neanderthal Y chromosome is compared with an average modern human Y chromosome using all linked

deleterious mutations on each simulated chromosome (14). This allows us to make a general statement about the probability of replacement in terms of the difference in fitness between Neanderthal and modern human Y chromosomes while abstracting

RESEARCH | REPORT

over other factors that affect reproductive fitness but are currently impossible to simulate accurately (26).

For example, assuming 5% gene flow from modern humans, we found that even a 1% reduction in Neanderthal Y chromosome fitness increases the probability of replacement after 50 ka to ~25%, and a 2% reduction in fitness increases this probability to ~50% (Fig. 3B). However, the rate of gene flow as well as any factor that contributes to the difference in fitness between Neanderthal and modern human Y chromosomes will have a pronounced effect on the replacement probability (figs. S27 to S32). Given the crucial role of the Y chromosome in reproduction and fertility and its haploid nature, it is possible that deleterious mutations or structural variants on the Y chromosome have a larger impact on fitness than considered in our simulations. We therefore refrain from making predictions about the specific process of replacement, because we lack information about the frequencies of introgressed Y chromosomes in older Neanderthals, potential sex bias in the gene flow, and the fitness effects of single-nucleotide and structural variants on the Y chromosome (26). Nevertheless, our models are a proof-of-principle demonstration that even a simple difference in the efficacy of purifying selection between two lineages can markedly affect introgression dynamics of nonrecombining, uniparental DNA.

We conclude that the Y chromosomes of late Neanderthals represent an extinct lineage closely related to modern human Y chromosomes that introgressed into Neanderthals between ~370 and ~100 ka ago. The presence of this Y chromosome lineage in all late Neanderthals makes it unlikely that genetic changes that accumulated in Neanderthal and modern human Y chromosomes before the introgression led to incompatibilities between these

groups (10). Furthermore, we predict that the ~400-ka-old Sima de los Huesos Neanderthals should carry a Y chromosome lineage more similar to that of Denisovans than to that of later Neanderthals (8, 9). Although the amount of modern human gene flow into Neanderthals appears to have been limited (13, 20, 21), we demonstrate that the replacement of mtDNA and Y chromosomes in Neanderthals is highly plausible, given the higher genetic load in Neanderthals compared with that in modern humans. Our results imply that differences in genetic load in uniparental loci between two hybridizing populations is a plausible driver for the replacements observed in other hybridization events (27–29).

REFERENCES AND NOTES

1. R. E. Green *et al.*, *Science* **328**, 710–722 (2010).
2. M. Meyer *et al.*, *Science* **338**, 222–226 (2012).
3. S. Lippold *et al.*, *Investig. Genet.* **5**, 13 (2014).
4. M. Karmin *et al.*, *Genome Res.* **25**, 459–466 (2015).
5. I. Olalde *et al.*, *Science* **363**, 1230–1234 (2019).
6. K. Prüfer *et al.*, *Nature* **505**, 43–49 (2014).
7. C. Posth *et al.*, *Nat. Commun.* **8**, 16046 (2017).
8. M. Meyer *et al.*, *Nature* **505**, 403–406 (2014).
9. M. Meyer *et al.*, *Nature* **531**, 504–507 (2016).
10. F. L. Mendez, G. D. Poznik, S. Castellano, C. D. Bustamante, *Am. J. Hum. Genet.* **98**, 728–734 (2016).
11. S. Sawyer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15696–15700 (2015).
12. K. Douka *et al.*, *Nature* **565**, 640–644 (2019).
13. M. Hajdinjak *et al.*, *Nature* **555**, 652–656 (2018).
14. See supplementary materials.
15. R. E. Wood *et al.*, *Archaeometry* **55**, 148–158 (2013).
16. S. Castellano *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6666–6671 (2014).
17. Q. Fu *et al.*, *Nature* **514**, 445–449 (2014).
18. S. Mallick *et al.*, *Nature* **538**, 201–206 (2016).
19. K. Prüfer *et al.*, *Science* **358**, 655–658 (2017).
20. M. Kuhlwilm *et al.*, *Nature* **530**, 429–433 (2016).
21. M. J. Hubisz, A. L. Williams, A. Siepel, *PLoS Genet.* **16**, e1008895 (2020).
22. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (The Blackburn Press, 1970).
23. K. Harris, R. Nielsen, *Genetics* **203**, 881–891 (2016).
24. I. Juric, S. Aeschbacher, G. Coop, *PLoS Genet.* **12**, e1006340 (2016).
25. B. C. Haller, P. W. Messer, *Mol. Biol. Evol.* **36**, 632–637 (2019).
26. S. Colaco, D. Modi, *Reprod. Biol. Endocrinol.* **16**, 14 (2018).
27. J. W. Ballard, M. C. Whitlock, *Mol. Ecol.* **13**, 729–744 (2004).
28. T. Bonnet, R. Leblois, F. Rousset, P.-A. Crochet, *Evolution* **71**, 2140–2158 (2017).
29. F. A. Seixas, P. Boursot, J. Melo-Ferreira, *Genome Biol.* **19**, 91 (2018).
30. L. Skov, M. H. Schierup; Danish Pan Genome Consortium, *PLoS Genet.* **13**, e1006834 (2017).
31. M. Petr, “The evolutionary history of Neanderthal and Denisovan Y chromosomes” - code and Jupyter notebooks. Zenodo (2020); doi:10.5281/zenodo.3941654.
32. M. Petr, “The evolutionary history of Neanderthal and Denisovan Y chromosomes” - Y chromosome capture designs. Zenodo (2020); doi: 10.5281/zenodo.3940568.

ACKNOWLEDGMENTS

We thank S. Pääbo, M. Stoneking, B. Peter, M. Slatkin, L. Skov, and E. Zavala for helpful discussions and comments on the manuscript. **Funding:** Q.F. was supported by funding from the Chinese Academy of Sciences (XDB26000000) and the National Natural Science Foundation of China (91731303, 41925009, 41630102). A.R. was funded by Spanish government (MICINN/FEDER) (grant number CGL2016-75109-P). The reassessment of the Spy collection by H.R., I.C., and P.S. was supported by the Belgian Science Policy Office (BELSPO 2004-2007, MO/36/0112). M.V.S., M.B.K., and A.P.D. were supported by the Russian Foundation for Basic Research (RFBR 17-29-04206). This study was funded by the Max Planck Society and the European Research Council (grant agreement number 694707). **Author contributions:** M.P. and J.K. analyzed data. M.H., Q.F., and E.E. performed laboratory experiments. H.R., I.C., P.S., L.V.G., V.B.D., C.L.-F., M.d.I.R., A.R., M.V.S., M.B.K., and A.P.D. provided samples. B.V., M.M., and J.K. supervised the project. M.P. and J.K. wrote and edited the manuscript with input from all co-authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Complete source code for data processing and simulations, as well as Jupyter notebooks with all analyses and results, can be found at Zenodo (31). Coordinates of the capture target regions and sequences of the capture probes are available at Zenodo (32). All sequence data are available from the European Nucleotide Archive under the accession number PRJEB39390.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6511/1653/suppl/DC1
Materials and Methods
Figs. S1 to S32
Tables S1 to S14
References (33–67)
MDAR Reproducibility Checklist

[View/request a protocol for this paper from Bio-protocol.](#)

9 March 2020; accepted 6 August 2020
10.1126/science.abb6460



science.sciencemag.org/content/369/6511/1653/suppl/DC1

Supplementary Materials for **The evolutionary history of Neanderthal and Denisovan Y chromosomes**

Martin Petr*, Mateja Hajdinjak, Qiaomei Fu, Elena Essel, H el ene Rougier,
Isabelle Crevecoeur, Patrick Semal, Liubov V. Golovanova, Vladimir B. Doronichev,
Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Michael V. Shunkov,
Maxim B. Kozlikin, Anatoli P. Derevianko, Benjamin Vernot,
Matthias Meyer, Janet Kelso*

*Corresponding author. Email: mp@bodkan.net (M.P.); kelso@eva.mpg.de (J.K.)

Published 25 September 2020, *Science* **369**, 1653 (2020)
DOI: [10.1126/science.abb6460](https://doi.org/10.1126/science.abb6460)

This PDF file includes:

Materials and Methods
Figs. S1 to S32
Tables S1 to S13
Caption for Table S14
References

Other Supplementary Material for this manuscript includes the following: (available at science.sciencemag.org/content/369/6511/1653/suppl/DC1)

MDAR Reproducibility Checklist (.pdf)
Tables S14 (.xlsx)

Y chromosome DNA capture design

To design a set of DNA capture probes, we identified regions of the human Y chromosome that are uniquely mappable with short sequence reads. Starting from the entire human Y chromosome reference sequence (version *hg19*), we removed regions that overlap those found by the Tandem Repeats Finder (33) and those identified by a previously described mappability track as regions that may result in ambiguous alignment of short reads (so called “*map35_50%*” filter, (6)). We then removed any regions that were shorter than 99 bp of continuous sequence. In total, this process yielded 6,912,728 bp (~6.9 Mb) of the Y chromosome suitable for use as an ancient DNA capture target.

We designed 52 bp oligonucleotide probes by tiling the identified 6.9 Mb of target sequence with 52 bp fragments in steps of 3 bp. This resulted in 2,049,846 individual oligonucleotide probes. To verify that the probe sequences are unique genome-wide, we aligned each probe to the complete *hg19* reference sequence and confirmed that they all aligned only to their expected position on the Y chromosome with mapping quality of at least 30. The files containing the coordinates of target regions, as well as the coordinates and sequences of all capture probes, including 8 bp adapters, are available for download at <https://doi.org/10.5281/zenodo.3940567>.

Following the approach taken by Fu *et al.* (34), 60 bp oligonucleotides containing the probe sequences as well as an 8 bp universal linker sequence were synthesized on three One Million Feature Arrays (Agilent Technologies), converted into probe libraries and amplified. Single-stranded biotinylated DNA probes were generated using a linear amplification reaction with a single biotinylated primer (34).

We also co-analyzed data from two additional captures carried out previously: (i) ~120 kb of Y chromosome sequence from the *El Sidrón 1253* Neanderthal that was targeted as a part of an exome capture study (16) and has been analyzed previously (10), and (ii) a larger amount of data (~560 kb) from the same *El Sidrón 1253* individual which we captured using probes designed for a previously published set of Y chromosome target regions (3). The files containing the coordinates of target regions and coordinates and sequences of all capture probes, including 8 bp adapters, are available for download at <https://doi.org/10.5281/zenodo.3940567>.

The features of our new capture design, as well as a comparison with the Y chromosome target regions on the exome capture (10, 16) and the ~560 kb capture (3) are reported in Table S1.

Comparison of our 6.9 Mb capture target to a previous Y chromosome capture design

In a study by Cruz-Dávalos *et al.* (35), the authors presented a different strategy for the enrichment of Y chromosomal hominin DNA. They started with 10.3 Mb of Y chromosome sequence defined as suitable for shotgun sequencing and genotype calling and used a proprietary algorithm to further restrict this sequence to regions usable for designing capture probes. This led to 8.9 Mb of capture sequence, which is approximately 2 Mb more than our capture target set. We note, however, that our starting set of regions for probe design was much more conservative and stricter.

In order to evaluate the differences between both capture sets, we downloaded the coordinates of the 8.9 Mb capture regions published by Cruz-Dávalos *et al.* (35) and applied filtering steps used to generate the target set in our study. First, we applied the 35-mer 50% mappability filter (6) designed to restrict to those regions of the genome suitable for mapping extremely short aDNA fragments. Interestingly, this leads to the removal of ~1.7 Mb out of the starting 8.9 Mb (approximately 20%). This step also leads to a large fragmentation of continuous regions (30,102 regions compared to 4,109 for the original 8.9 Mb starting set), many of them only 1 bp long. When we applied our second filtering criterion on top of the mappability filter, removing all regions shorter than 99 bp, we were left with 6.4 Mb regions (down from the original 8.9 Mb; loss of ~28% of sequence from the original 8.9 Mb). This is remarkably close to our own capture target of 6.9 Mb. In fact, the overlap between the strictly filtered set of the original 8.9 Mb and our own capture target is ~6.3 Mb (90%) and the distributions of lengths of both sets of capture regions is also remarkably similar (Fig. S1). This shows that most of the differences between our capture target and that of Cruz-Dávalos *et al.* (35) lies in the specific design of our capture set to target extremely short archaic human DNA fragments.

We also note that another significant factor is the length of probe sequences and probe tiling density, both of which will play a significant role in the efficiency of DNA capture (34). The impact of probe tiling density on the efficiency of capture of archaic and modern human Y chromosomes remains to be determined.

Sampling, DNA extraction, library preparation and capture

Samples of 15.4 mg and 14.9 mg of tooth powder from *Denisova 8* were used for DNA extraction using a silica-based method (36) with modifications as described in (37). Ten mg of the tooth powder from *Denisova 4* were used for a silica-based DNA extraction that is optimized for the recovery of extremely short DNA fragments (38). Four samples of Mezmaiskaya 2 bone powder, ranging between 3.2 mg and 17.5 mg were treated with 0.5% hypochlorite solution to minimize microbial and present-day human DNA contamination (37) before DNA was extracted either manually (36) or on an automated liquid handling platform (Bravo NGS workstation B, Agilent Technologies) (39). See Table S14 for an overview of the DNA extracts and libraries generated in this and previous studies and the experimental conditions used.

In addition to existing single-stranded libraries for *Spy94a* and *Mezmaiskaya 2* (13), new single-stranded DNA libraries for *Mezmaiskaya 2*, *Denisova 4* and *Denisova 8* were prepared from DNA extracts made for this study (Table S14). Two of the single-stranded DNA libraries for *Denisova 8* (A9461 and A9462) were prepared manually using 10 μ L of each extract as an input (40). All other single-stranded DNA libraries were prepared using either 10 μ L or 30 μ L of extract as input (41) on an automated liquid handling platform (Bravo NGS workstation B, Agilent Technologies) (42). All new libraries were prepared without UDG treatment (non-UDG treated libraries).

In order to monitor the efficiency of library preparation, a control oligonucleotide was spiked into each aliquot of a DNA extract used for library preparation (38). Quantitative PCR was used to determine the total number of unique library molecules and

the number of oligonucleotides that were successfully converted to library molecules (38, 41) (Table S14). Each library was tagged with two unique index sequences (43) and amplified into plateau with AccuPrime Pfx DNA polymerase (Life Technologies) (44) according to the modifications detailed in (37). Fifty microlitres (half of the total volume) of each of the amplified libraries were purified on an automated liquid handling platform (Bravo NGS workstation B, Agilent Technologies) using SPRI beads (42). A NanoDrop 1000 Spectrophotometer (NanoDrop Technologies) was used to determine the concentrations of the purified libraries.

In solution hybridization capture of the Y chromosome was performed in two successive rounds of capture as described previously (34), using the Y chromosome probe set designed in the present study and single-stranded libraries prepared in this and previous studies. In addition, we performed hybridization capture on 40 double-stranded libraries prepared in a previous study from the *El Sidrón 1253* Neanderthal (see Table S1 in (16)) using a smaller ~560 kb Y chromosomal probe set that was also designed previously (3).

Although previous studies identified other male archaic human individuals such as Devil's Tower and Sima de los Huesos, their DNA preservation turned to be too poor to be considered for Y chromosome capture (9, 45).

Sequencing and data processing

Newly generated archaic human Y chromosomes

All captured libraries were sequenced on the Illumina HiSeq 2500 platform in a double index configuration (2x76 cycles) (43), and base calling was done using Bustard (Illumina).

Adapters were trimmed and overlapping paired-end reads were merged using a Bayesian approach implemented in *leeHom* using the following command-line: `leeHom -f <adapter1> -s <adapter2> -c <possible chimeras> --ancientdna (46)`. The Burrows-Wheeler Aligner (BWA) (47) with parameters adjusted for alignment of ancient DNA (“-n 0.01 -o 2 -l 16500”) was used to align the sequenced fragments to the human reference genome version hg19/GRCh37. Only reads showing perfect matches to the expected index sequence combinations were retained for subsequent analyses. PCR duplicates were removed using the *bam-rmdup* program, which can be downloaded in source form from <https://github.com/mpieva/biohazard-tools>. DNA fragments that were at least 35 base pairs (bp) long and had a mapping quality of at least 25 were extracted using *samtools* (48). Each processed and filtered BAM file (one for each archaic human Y chromosome) was intersected with a BED file of the appropriate Y chromosome target (full ~6.9 Mb capture, ~120 kb exome capture or ~560 kb capture).

To facilitate further research as more human genomic data transitions to the newest GRCh38 reference genome, we also provide our 6.9 Mb archaic human Y chromosome capture data mapped to GRCh38 using the same parameters as the hg19/GRCh37 mapping. The sequences filtered for length ≥ 35 bp and mapping quality ≥ 25 are available as BAM files at <https://doi.org/10.5281/zenodo.3946496>.

Previously published archaic human sequences

In addition to the new capture data generated here, we analyzed previously published shotgun sequences of the *Spy 94a* and *Mezmaiskaya 2* individuals (13), as well as exome capture data of the *El Sidrón 1253* individual (16). For comparisons to our capture data, we

generated BAM files for *Spy 94a* and *Mezmaiskaya 2* shotgun sequences and the *El Sidrón 1253* exome capture by filtering the published data to minimum read length of 35 bp and mapping quality 25, keeping only sequences aligned to the set of appropriate target capture regions (~6.9 Mb capture target for *Spy 94a* and *Mezmaiskaya 2*, ~118 kb capture target for *El Sidrón 1253*, Table S1).

Previously published modern human sequences

For comparisons with modern human Y chromosomes, we downloaded BAM files of African and non-African Y chromosomes published by the Simons Genome Diversity Project (SGDP) (18), two Y chromosomes representing the African A00 lineage (4) and the Y chromosome of a ~45,000-year-old hunter-gatherer *Ust'-Ishim* (17) (Tables S4 and S10). Because the two individuals from which the A00 Y chromosomes were sequenced are closely related and each is only about half of the coverage of the other modern human Y chromosomes (Table S4), we followed the approach of the original A00 publication and merged the two A00 Y chromosomes into a single BAM file (4, 49). All individual BAM files (one for each modern human Y chromosome) were then filtered to retain reads with a minimum length of 35 bp and mapping quality of at least 25, and alignment to the appropriate set of Y chromosome target capture regions (Table S1).

Coverage and measures of ancient DNA quality

Coverage

Sequencing coverage was calculated using *bedtools* (50). To get coverage for a given individual in a given set of target regions, we ran the command `bedtools coverage`

-a <BED> -b <BAM> -d, which reports the coverage for each position in a BED file in the last column of its output. We removed sites with coverage higher than the 98% quantile of the entire distribution in each of the individuals in our study. Fig. 1C (spatial distribution and overall distribution) and Tables S2, S3 and S4 summarize the values of coverage at sites with less than 98% quantile of the overall distribution in a sample.

Patterns of ancient DNA damage

To check for the presence of genuine ancient DNA sequences, we looked for an increased rate of deamination-induced substitutions, an important signature of ancient DNA damage (51). We counted substitution frequencies for each individual BAM file (one BAM file per individual Y chromosome) and found that molecules from single-stranded libraries that were not treated by uracil-DNA glycosylase (UDG) enzyme (those from *Spy 94a*, *Mezmaiskaya 2*, *Denisova 4* and *Denisova 8*) show highly elevated frequencies of C-to-T substitutions towards the ends of molecules, as well as C-to-T substitutions throughout the molecules (Fig. S2). As is characteristic of double-stranded libraries treated with the UDG enzyme, the deamination substitution frequency signal in the capture data from *El Sidrón 1253* UDG-treated libraries is much less pronounced and present only at the terminal positions of DNA fragments as both C-to-T and G-to-A substitutions (Fig. S3). For comparison, Fig. S4 shows DNA damage patterns from previously published shotgun sequences of *Spy 94a* and *Mezmaiskaya 2* individuals.

Read length distribution

We calculated read lengths for each final processed BAM file using `samtools view` and `awk`. As expected for ancient sequences, archaic human Y chromosome fragments are very short (Fig. S5, Table S5). We note that *Denisova 8* shows an even more extreme reduction in read length compared to the other captured archaic human Y chromosomes (Fig. S5, Table S5), consistent with the fact that the *Denisova 8* specimen is possibly nearly twice as old as the other archaic humans in our study (Fig. 1A).

Capture enrichment efficiency

We assessed the performance of the Y-chromosome capture by comparison of the on-target fraction in the captures to the on-target fraction in shotgun sequencing data generated from the same libraries. For all libraries the fold enrichment is estimated to be between 300- and 1500-fold (Table S14).

Modern human contamination

Our consensus-based genotype calling strategy is designed to remove the effect of modern human contamination under the assumption that contaminant reads at a given position never represent more than 90% of the total number of reads. To validate that this approach achieves the desired effect, we assessed the frequency of modern human-derived SNPs at positions informative about modern human contamination in the final archaic human Y chromosome genotype calls.

To define these informative positions, we used genotypes of present-day human Y chromosomes and identified ancestral states by determining which sites carry the same

allele in two present-day African lineages *A00* and *S_Ju_hoan_North-1* (4, 18) (red branches in Fig. S6). We then further restricted these sites to those at which a different allele is observed in all 13 non-African individuals from the SGDP panel (18). These represent alleles derived on the non-African Y chromosome lineage (blue branches in Fig. S6). This conditioning led to a total of 268 informative positions. Given that all archaic human Y chromosomes are expected to carry the ancestral state at these sites because they all fall basal to modern human Y chromosomes (Fig. 2A), observing a derived allele at any of these informative sites implies the presence of a modern human contaminant allele, double mutation or an erroneous SNP call. We note that although the 13 non-African Y chromosomes that we used to define the potential ‘contaminant-derived states’ may not represent the true contaminant population, the contaminating population would still share the same derived states due to the non-recombining nature of human Y chromosomes.

Using this set of 268 informative positions, we found that the five archaic human Y chromosomes carry the ancestral state at all informative positions except for a single position in the *Spy 94a* individual which shows a derived allele out of the total 16 informative sites available (Table S6). This shows that the consensus genotype calling method is efficient in mitigating the effect of modern human contaminant reads on the final set of Y chromosome genotype calls.

To further validate these conclusions, we repeated our main analysis (TMRCA estimation between archaic and modern human Y chromosomes) on reads showing evidence for aDNA deamination (a signature of truly ancient sequences) at the terminal three positions of the molecules. In agreement with the contamination analysis described in this section,

we did not find any difference between results obtained on the two versions of the data which demonstrates that contamination does not affect our conclusions (Fig. S24).

Capture bias and reference (mapping) bias

Because the probes we designed for Y chromosome DNA enrichment are based on the human reference genome sequence, we were concerned about the effect of capture bias on our inferences, specifically on the observed differences in divergence times between Denisovan and Neanderthal Y chromosomes with respect to present-day humans (Fig. 2). An earlier study of reference bias in published aDNA data sets has found minor but significant allelic imbalances at heterozygous sites from a baseline expectation of 50% ratio between reference and alternative alleles (52). Because this approach is not applicable for haploid Y chromosomes, we instead looked for departures of the observed number of sites without any genomic coverage from the theoretical expectation.

To build an intuition about this expectation, let's first consider a case of a truly random distribution of sequencing reads in a complete absence of capture or reference bias. In such a situation, the count of reads observed at any site can be modeled as a random variable which follows a Poisson distribution with a parameter λ , where λ represents the average coverage observed across all sites. In a mathematical notation, letting X be this count of reads: $X \sim \text{Poisson}(\lambda)$. Then, given some value of λ , the expected proportion of sites that are not covered by any sequencing reads can be expressed as $\text{Poisson}(X = 0, \lambda)$, i.e. as the probability of observing zero reads at any site given the overall average coverage of λ . As an example, assuming 1-fold sequencing coverage we would expect to see $\text{Poisson}(X = 0, \lambda = 1) = 0.3678794 \sim 37\%$ of the target sites not to be covered by any

read at all, just by random chance. Importantly, however, capture bias or reference bias will manifest by some regions of the genome being underrepresented in terms of captured molecules or mapped reads. Therefore, the presence and magnitude of this bias in a given DNA enrichment experiment can be detected by estimating the difference between the proportion of sites without any sequencing coverage from the theoretical Poisson expectation.

The results, shown in Fig. S7 and Table S7, demonstrate that there is both reference and capture bias in our data and offer several interesting insights. First, we see a comparable effect of bias in all capture data (4-6% departure from the theoretical Poisson expectation) regardless of which capture array was used for the enrichment procedure (i.e., the full 6.9 Mb capture array, the 560 kb capture array or the exome capture array, Fig. S7). Furthermore, comparisons of capture and shotgun sequences of *Spy 94a* and *Mezmaiskaya 2* show that the majority of bias must be due to the capture procedure itself (failure to capture molecules). This is because the underlying true biological divergences of *Spy 94a* and *Mezmaiskaya 2* to the reference genome (which cause a failure to map reads due to an increased number of substitutions – i.e., a reference bias) must be the same for both capture and shotgun sequences from these individuals. Crucially, however, despite the differences in bias between capture and shotgun sequences, both datasets lead to the same estimates of TMRCA with present-day human Y chromosomes (Fig. S21). Furthermore, although we see dramatically different phylogenetic relationships of Denisovan and Neanderthal Y chromosomes with respect to modern humans (Fig. 2A), both groups of archaic human capture sequences display comparable magnitudes of both sources of bias (Fig. S7).

Genotype calling

Consensus genotype calling

The haploid nature of the human Y chromosome alleviates many issues inherent to genotype calling of diploid genomes. Most importantly, given that only one allele is expected to be present at each site of a non-recombining portion of the Y chromosome, observing more than one allele at a site must be the result of sequencing errors, DNA damage, contamination, or misalignment of reads. While such issues present a significant problem for calling diploid genotypes, by making it challenging to distinguish true heterozygous calls from erroneously called heterozygous genotypes (53, 54), they are less of an issue for haploid genotyping.

To call genotypes of the archaic and modern human Y chromosomes in our study, we applied a conservative approach to produce a consensus of sequencing reads. For each Y chromosome BAM file, we performed a pileup of reads at each site (disabling base quality recalibration), filtering out reads with mapping quality less than 25, ignoring bases with base quality less than 20 and removing reads carrying indels at a pileup position. Then, under the assumption that alleles introduced due to DNA damage, sequencing errors, misalignments, or contamination will be in a minority at each site, we called the allele supported by at least 90% of the reads in a pileup as the haploid genotype for that site. For further analyses, we additionally restricted to genotype calls supported by at least three reads. This genotype calling procedure has been implemented in a Python program which is available in our project GitHub repository at <https://github.com/bodkan/archaic-ychr>. In addition, a Zenodo archive with the version of the code used here is available at <https://doi.org/10.5281/zenodo.3890512>.

Genotype calling using *snpAD*

The consensus genotype calling approach described in the previous section is quite conservative and does not incorporate an explicit model of DNA damage and sequencing errors. To validate the robustness of our consensus-based results, we compared them to genotype calls generated using *snpAD*, an aDNA-specific genotype caller (54). A major caveat of this approach is the fact that *snpAD* has been designed for calling diploid genotypes and accurate results requires at least 4X genomic coverage (54). Therefore, its genotype calling model has not been tested on low coverage, haploid chromosomes such as those generated in our study. While recognizing these limitations, we used *snpAD* to call genotypes of all four archaic Y chromosomes captured for the 6.9 Mb target regions (*Denisova 4*, *Denisova 8*, *Spy 94a* and *Mezmaiskaya 2*), discarded any sites which were called as heterozygous (likely the result of errors, contamination or aDNA damage), and converted all homozygous genotype calls to a haploid state.

In accordance with *snpAD*'s more sophisticated model of aDNA damage patterns, we have found that the number of successfully genotyped sites is higher than those generated by our simpler consensus-based genotype calling approach, but only marginally so (Table S8). Furthermore, although the rates of C-to-T and G-to-A SNP frequencies observed in the final set of genotype calls of the high coverage *Mezmaiskaya 2* is very close to the baseline expectation for present-day DNA, the remaining low coverage archaic Y chromosomes are still affected by aDNA damage and show an excess of falsely called genotypes (Figs. S8 and S9). We note that this is not unexpected, because the coverage of these individuals is much lower than what is recommended for *snpAD* (54). Overall, we did not observe significant differences between *snpAD*-based and consensus-based

genotypes in terms of the inferred times to the most recent common ancestor (TMRCA) and, in fact, we found that both lead to the same conclusions (Fig. S21). Based on these analyses we concluded that our conservative 90% cutoff for consensus genotype calling method is appropriate and decided to use it for all analyses.

Minimum coverage filtering

The majority of libraries analyzed in our study have not been treated with the uracil-deglycosylase (UDG) enzyme. Unlike UDG-treated libraries, non-UDG libraries retain an increased deamination signal throughout the molecules (Figs. S2 and S3) which poses a significant challenge for distinguishing false substitutions caused by aDNA damage from true polymorphisms (54).

For a given sequencing read carrying a putative substitution, it is not straightforward to decide whether this substitution represents a true polymorphism or error. Given enough sequencing coverage, this issue can be mostly overcome by observing a sufficient number of bases from reads that do not carry a deamination-induced substitution, integrating evidence from multiple reads at a site (54). However, as our data is of relatively low coverage (Fig. 1C, Table S2), we were concerned by selecting an appropriate lower coverage cutoff to minimize the impact of false polymorphisms on our inferences. Specifically, if the same nucleotide is observed in a majority of reads mapped to the same genomic position, it is unlikely that this would be the result of aDNA damage, sequencing errors or contamination, as these occur mostly at relatively low frequencies in the individuals in our study (Fig. S2 and Table S6). To get a sense of the frequency of calling false polymorphisms as a function of coverage, we calculated the proportions of observed genotypes in each archaic human Y chromosome given a certain coverage filtering cutoff

and compared those to the baseline expectation for present-day human DNA. As expected, allowing SNPs supported by only one read leads to a significant excess of C-to-T and G-to-A SNP (Fig. S8), a consequence of the presence of aDNA damage (Fig. S2). We found that increasing the minimum coverage cutoff to two reads causes the rate of aDNA-induced SNPs to drop significantly towards the baseline expectation, but going beyond requiring the support of three reads for each SNP does not lead to further improvement in accuracy of genotype calling (Fig. S8). Because of this and because each additional increase in required minimum coverage is at the expense of the final number of available sites, we settled on a minimum coverage cutoff of 3 reads.

It is important to note that despite the residual presence of false aDNA substitutions in the final set of filtered genotype calls, manifesting as increased frequencies of C-to-T and G-to-A SNPs compared to present-day DNA (Fig. S8), comparisons of archaic-modern human TMRCA estimates obtained using the full set of genotype calls and those based on genotypes restricted to non-C-to-T/G-to-A SNPs did not reveal any significant differences (Fig. S19). This is partially due to very low rates of residual false SNPs that pass through the filtering, but mostly because our TMRCA estimators are quite insensitive to private mutations on the archaic lineage. A second validation of our coverage filter follows from the fact that the two Denisovan and all three Neanderthal Y chromosomes lead to the same TMRCA estimates with modern human Y chromosomes despite differences in coverage and rates of aDNA damage (Figs. 1C, 2B and S8). Both of these factors would affect genotyping accuracy if not handled appropriately, and would introduce noise in TMRCA estimates estimated for individual Y chromosomes.

Complete counts of Y chromosome positions passing the different filters applied cumulatively (all callable sites, sites covered by at least three reads and sites with coverage less than the 98% quantile of the coverage distribution) are reported in Table S8 (6.9 Mb target regions) and Table S9 (560 kb target regions).

Inferring phylogenetic relationships

To resolve the phylogenetic relationships of each archaic human Y chromosome to all other Y chromosome sequences, we merged the VCF files with genotype calls from each individual (including the genotypes of the chimpanzee reference genome (panTro4)) into a single VCF file and converted the genotypes to the FASTA format using a custom Python script (available on our GitHub repository: <https://github.com/bodkan/archaic-ychr> and in the Zenodo archive at <https://doi.org/10.5281/zenodo.3890512>). To mitigate biases introduced by low coverage and characteristics of aDNA damage (55), we excluded all C-to-T and G-to-A polymorphisms and applied the same filters for each individual as for all other analyses in our study. Finally, we excluded monomorphic sites and sites carrying private changes on the chimpanzee lineage to reduce the size of the final alignment file.

To construct a neighbor-joining phylogenetic tree (Fig. 2A), we utilized the functionality provided by R packages *ape* and *phangorn* (56, 57). First, we calculated the distance matrix between all Y chromosome pairs in the FASTA file with the function `dist.dna`, using the model of simple pairwise differences (`model = "raw"`) and excluding sites with missing data specific to each pair (`pairwise.deletion = TRUE`). We then provided this distance matrix to the `nj` function and rooted the resulting neighbor-joining tree using the function `midpoint` from the *phangorn* package. Bootstrap confidence numbers for the neighbor-joining tree were calculated using *ape*'s `boot.phylo` function over 100 replicates. After inspecting the resulting phylogenetic tree, we found that the private branch leading to the *Denisova 4* had a negative length (value = -0.00088). Given that negative branch lengths are a relatively common artefact of the neighbor-joining algorithm and do not affect the reliability of the generated tree we followed the

recommendation to set the branch length to zero (58). We note that this does not have any impact on our conclusions, because the change involves a private branch whose length is not meaningful given the discrepancies between sample dates and implied tree tip dates (Figs. 1A and 2A, (53)). The final trees were annotated and plotted using the R package *ggtree* (59).

Estimating the TMRCA of archaic and modern human Y chromosomes

Given that most of the Y chromosome capture data analyzed in our study is of relatively low coverage (Fig. 1C, Table S2), care needs to be taken when estimating phylogenetic parameters such as the time to the most recent common ancestor (TMRCA). Similarly, low coverage and the associated reduction in the accuracy of genotype calls render the inferred aDNA branch lengths unreliable (55). Any phylogenetic method of choice must be therefore robust to sequencing errors and incorrect branch lengths. We also observe discordances between sample dates and implied molecular tip dates, likely due to residual genotype calling errors (compare Fig. 1A vs Fig. 2A). We therefore estimated TMRCA between archaic and modern human Y chromosomes using a method inspired by the analysis of the *El Sidrón 1253* Neanderthal coding sequence (10). Instead of using polymorphisms on the archaic human lineage, this method relies on first estimating the TMRCA of a pair of high-coverage African and non-African Y chromosomes ($TMRCA_{AFR}$) which is then used to extrapolate the deeper divergence time between archaic and modern

human Y chromosomes. We describe the method in the sections below, detailing our modifications and improvements.

TMRCAs of Africans and non-Africans ($TMRCAs_{AFR}$)

The original study of the *El Sidrón 1253* Neanderthal Y chromosome estimated the TMRCAs between the A00 African Y chromosome lineage and the hg19 Y chromosome as a representative of non-African Y chromosomes (4, 10). In order to get a better sense of the uncertainty and noise in our TMRCAs estimates, we expanded the present-day Y chromosome reference panel to 13 non-African and 7 African Y chromosomes from the SGDP data set (Table S10) (18).

In the first step, we estimated mutation rate in the 6.9 Mb capture target using the high-coverage Y chromosome of *Ust'-Ishim*, a 45,000 years old hunter-gatherer from Siberia (17). We counted derived mutations missing on the *Ust'-Ishim* branch compared to those observed in the panel of 13 non-African Y chromosomes, and used this branch-shortening to calculate mutation rate (Fig. S11, Table S11). In the second step, we counted mutations accumulated on an African lineage and a non-African lineage since their split from each other and calculated the TMRCAs of both (in units of years ago) using the mutation rate estimated in the first step. Importantly, we discovered that the branch-lengths in Africans are as much as 13% shorter compared to non-Africans (Fig. S13), which is consistent with significant branch length variability discovered in previous studies and suggested to be a result of various demographic and selection processes (60, 61). To keep our methodology consistent throughout our analyses, we estimated the TMRCAs of African and non-African Y chromosomes ($TMRCAs_{AFR}$) as the length of the non-African Y lineage

(sum of branch lengths $a + d$ in Fig. S11). Encouragingly, we found that our mutation rate and $TMRCA_{AFR}$ point estimates (Table S11) match closely those based on a large panel of present-day Y chromosomes (4). Most importantly, using the A00 lineage as a representative of the deepest known split among present-day human Y chromosomes we inferred a $TMRCA_{A00}$ of ~249 years ago (point estimate based on an estimated mutation rate of 7.34×10^{-10} per bp per year), which is comparable to the $TMRCA_{A00}$ of ~254 years ago estimated by Karmin *et al.* (4). Therefore, our more restricted 6.9 Mb capture target gives TMRCA estimates consistent with those obtained from the full Y chromosome shotgun data (4).

Archaic human-modern human TMRCA ($TMRCA_{archaic}$)

Having estimated $TMRCA_{AFR}$ (described in the previous section), we can express the TMRCA of archaic and modern human Y chromosomes ($TMRCA_{archaic}$) as a factor of how much older is $TMRCA_{archaic}$ compared to $TMRCA_{AFR}$ (Fig. S14). In mathematical terms, if we call the scaling factor α (following the terminology of (10)), we can write

$$TMRCA_{archaic} = \alpha \times TMRCA_{AFR} . \quad (1)$$

In the remainder of this section, we present two ways of calculating α , first using the original approach of Mendez *et al.* (10) and then using a more straightforward method.

The previous approach to calculate α

Based on Fig. S14, an alternative way to express $TMRC A_{archaic}$ in addition to equation (1) is

$$TMRC A_{archaic} = T_{shared} + TMRC A_{AFR}. \quad (2)$$

The expressions (1) and (2) define a system of two equations and three variables, which can be solved for T_{shared} to get

$$T_{shared} = \alpha \times TMRC A_{AFR} - TMRC A_{AFR} = TMRC A_{AFR} \times (\alpha - 1). \quad (3)$$

Mendez *et al.* (10) found an expression for α by considering a ratio of time shared by hg19 and A00 Y chromosomes after their split from the *El Sidrón 1253* Neanderthal (T_{shared} above) and private branch lengths of both (Fig. S14), arriving at the following expression:

$$\frac{T_a}{T_a + T_d + T_e} = \frac{T_{shared}}{T_{shared} + 2 T_{AFR}} = \frac{T_{AFR} (\alpha - 1)}{T_{AFR} (\alpha - 1) + 2 T_{AFR}} = \frac{\alpha - 1}{\alpha + 1}. \quad (4)$$

Assuming mutation rate constancy on different lineages, α can be found by solving the following equation

$$\frac{a}{a + d + e} = \frac{\alpha - 1}{\alpha + 1},$$

which leads to

$$\alpha = \frac{2 a + d + e}{d + e}. \quad (5)$$

Using this expression for α and the values of $TMRC A_{AFR}$ estimated in the previous step, we can calculate $TMRC A_{archaic}$ for each pair of archaic and non-African Y chromosomes using equation (1) (Fig. S19B).

A more robust α statistic

While investigating the effect of minimum coverage filtering on genotype calling accuracy, we discovered a concerning dependence of the apparent branch lengths on the choice of the minimum coverage cutoff. Under normal conditions, the relative proportions of branch lengths a , d and e (Fig. S14) should remain constant regardless of coverage. This is crucial because the α estimator proposed by Mendez *et al.* (10) relies on proportions of lengths of all three of these branches (equations (4) and (5)).

Strikingly, we found that although the proportions of a and d branch lengths remain relatively stable even for extremely strict coverage filters, the relative length of the e branch (given by the proportion of derived mutations on the private African branch) has increasing tendency (Figs. S16 and S17). Furthermore, although this effect is most pronounced in low coverage samples (Figs. S16 and S17), it is clearly present even in the high coverage *Mezmaiskaya 2*, although at much higher coverage cutoffs (Fig. S18). Therefore, the issue is clearly not sample-specific but is a common artifact caused by pushing the minimum required coverage close to, or even beyond, the average coverage. Restricting to sites with a high number of aligned reads leads to enrichment of regions of lower divergence from the reference sequence, distorting the normal proportions of derived mutations observed on different branches of the tree.

In light of these issues, we note that there is a more straightforward way to express the scaling factor α :

$$\alpha = \frac{a + d}{d}. \quad (6)$$

This follows trivially from the definition of α as the factor of how much deeper $TMRCA_{archaic}$ is compared to $TMRCA_{AFR}$ and, unlike the original formulation of α (equation (5)), has the advantage of not relying on the relative length of the African branch e . This is important not only because of discordant branch proportion patterns (Figs. S16-S18) but also due to known unequal branch lengths observed in African and non-African Y chromosome lineages (Fig. S13, (60, 61)).

For completeness, we note that in an idealized situation without any coverage-dependent branch proportion bias we could assume $e \approx d$. Substituting for e in equation (5) would then give

$$\alpha = \frac{2a + d + e}{d + e} \approx \frac{2a + 2d}{2d} \approx \frac{a + d}{d}.$$

Therefore, under ideal conditions, both approaches to calculate α (equations (5) and (6)) are mathematically equivalent. However, in light of the coverage-dependent branch proportion bias shown in Figs. S16-S18, we calculate the α factor using equation (6), disregarding the branch length e completely.

Using the new expression for α and the values of $TMRCA_{AFR}$, we can estimate $TMRCA_{archaic}$ for each pair of archaic and non-African Y chromosomes using equation (1) (Fig. 2A, Table S12). By comparing TMRCA results based on the two formulations for α , we found similar estimates for most of the archaic human Y chromosomes in our study (Figs. S19 and S20). The only exception are the two Denisovan Y chromosomes, for which we infer slightly higher TMRCA with modern humans using the new α estimation

procedure compared to the formulation based on the original method (Figs. S19 and S20). This is a consequence of an increased proportion of the e branch relative to the d branch in *Denisova 4* and *Denisova 8* at the chosen minimum coverage filter which is evident in Fig. S17. Because the original method of Mendez *et al.* (10) relies on the e count of the derived African alleles (equations (4) and (5)), this leads to a slight decrease in the value of the α factor and, consequently, to a lower inferred TMRCA value. In contrast, our new formulation of α (equation (6)) is robust to this artifact and the inferred values of TMRCA are not affected.

Finally, we want to emphasize that although the analyses of branch length discrepancies discussed in this section were mostly based on results obtained using the A00 Y chromosome lineage, the issues we discovered are not specific to a particular choice of an African Y chromosome (Fig. S18A-C). However, comparisons of TMRCA estimates for the low coverage samples with those obtained for the high coverage samples (which do not show any biases at coverage cutoffs used throughout our study) clearly show that the inferences are most stable when the A00 lineage is used in the calculation of the α scaling factor, even for the samples with lowest coverage (Fig. S23). Therefore, all main results in our study are based on calculations using the A00 high coverage Y chromosome.

TMRCA of *Mezmaiskaya 2* and *Spy 94a*

The split time of Neanderthal and modern human Y chromosomes estimated in the previous section provides an upper bound for the last time the two populations experienced gene flow. Similarly, the deepest divergence in late Neanderthal Y chromosomes represents a lower bound, as the introgressed Y chromosome lineage must have already been present in Neanderthals prior to this diversification.

To estimate the deepest TMRCA of the known Neanderthal Y chromosomes (i.e. the TMRCA of *Mezmaiskaya 2* and *Spy 94a*, Fig. 2A), we first defined a set of sites in the ~6.9 Mb capture target regions which carry a reference allele in the chimpanzee, A00 and French Y chromosomes (the ancestral state) and an alternative allele (the derived state) on the branch leading to the high-coverage Y chromosome of *Mezmaiskaya 2* (Fig. S25), using the standard filtering used in previous sections (minimum three reads covering each genotyped site). We can calculate the approximate length of this branch using the TMRCA of *Mezmaiskaya 2* and modern human Y chromosomes (~370 kya, Table S12) and the known age of *Mezmaiskaya 2* (~44 kya, (13)) as 370 kya – 44 kya = 326 ky (Fig. S25). We can then estimate the split time between *Mezmaiskaya 2* and *Spy 94a* Y chromosomes using the proportion of *Mezmaiskaya*-derived sites which show the ancestral allele in *Spy 94a* (Fig. S25). Specifically, if we let Y be the number of ancestral alleles observed in *Spy 94a* and $X + Y$ be the total number of sites with genotype calls in *Spy 94a* at positions derived in *Mezmaiskaya 2*, we can express the TMRCA of Y chromosomes of the two Neanderthals simply as

$$\frac{Y}{X+Y} \times 326 + 44 \text{ kya} \quad (7)$$

(Fig. S25). We maximized the amount of data available for the analysis by merging the capture data with previously published shotgun sequences (13) and evaluated the robustness of the results to different genotype filtering and classes of polymorphisms. To estimate confidence intervals (C.I.), we re-sampled the X and Y counts from Poisson distributions with expected values given by the observed counts (Fig. S25), calculated the TMRCA on the re-sampled counts using equation (7) and then took the 2.5% and 97.5% quantiles of this simulated TMRCA distribution to arrive at the approximate range of 95% C.I.

The TMRCA of *Mezmaiskaya 2* and *Spy 94a* are consistently around ~100 kya regardless of the filtering criteria used (individual point estimates and 95% confidence intervals shown in Fig. S26 and Table S13). Together with the TMRCA of Neanderthal and modern human Y chromosomes, this suggests that the gene flow from an early population related to modern humans is likely to have happened some time between ~100 kya and ~370kya. We note that this time window is significantly wider than the one inferred based on a much more extensive set of available Neanderthal mtDNA genomes (219-468 kya) (7). However, it is likely that future sampling of Neanderthal Y chromosome diversity will reveal more basal Y chromosome lineages as has been the case for Neanderthal mtDNA (7).

Confidence intervals

Under the assumption that mutations on each branch of a tree (Figs. S11 and S14) accumulate independently, the observed counts of mutations can be modeled as realizations of independent Poisson processes (mutation counts in Tables S11 and S12). To quantify the uncertainty in our TMRCA estimates, we used a simulation-based bootstrapping approach. For each set of branch lengths used to calculate scaling factor α (branches a , d and e), we generated 1000 sets of simulated counts by randomly sampling from a Poisson distribution with the parameter λ set to values observed from the data. In other words, we simulated “trees” implicitly by generating a set of Poisson-distributed branch lengths. We then used the simulated counts to estimate the corresponding TMRCA values, obtaining a distribution of TMRCA consistent with the observed data. Finally, we took the lower 2.5% and upper 97.5% quantiles of the simulated distribution as the boundaries of bootstrap-based 95% confidence intervals.

To estimate the confidence interval for TMRCA across the whole panel of 13 non-African Y chromosomes (black dotted horizontal lines in TMRCA figures in our study such as Fig. 2B), we followed the same procedure but pooled all simulated counts together (i.e., 1000 simulated counts for each of the 13 Y chromosomes). Then, we took the lower 2.5% and upper 97.5% quantiles of TMRCA estimates calculated from the pooled counts.

Simulations of introgression under purifying selection

To investigate the expected frequency trajectories of Y chromosomes introgressed from modern humans into Neanderthals and the probability of their fixation, we adapted a modeling approach previously used to study negative selection against Neanderthal DNA in modern humans (23, 63). Briefly, this model assumes lower effective population size (N_e) in Neanderthals than modern humans as inferred from comparisons of whole-genome sequences (6). Under nearly-neutral theory, such differences in N_e are expected to increase the genetic load in Neanderthals compared to modern humans through an excess of accumulated deleterious mutations due to lower efficacy of purifying selection. Therefore, after introgression from Neanderthals into modern humans, Neanderthal haplotypes would be under stronger negative selection compared to modern human haplotypes, causing a rapid decrease in the proportion of genome-wide Neanderthal ancestry (23, 63).

In the context of evidence for Neanderthal Y chromosome replacement in our study, we were particularly interested in the dynamics of introgression in the opposite direction, from modern humans into Neanderthals. Specifically, given that nearly-neutral theory predicts that Neanderthal Y chromosomes would carry a higher load of deleterious mutations compared to modern human Y chromosomes, how much is natural selection expected to favor introgressed modern human Y chromosomes compared to their original Neanderthal counterparts?

Simulation setup

To address this question, we used a forward population genetic simulation framework SLiM (version 3.3) (25) to build an approximate model of modern human and Neanderthal demographic histories, following a strategy we used to study the long-term negative selection against Neanderthal autosomal DNA in modern humans (63). To simulate differences in N_e of both populations and to explore their impact on model dynamics, we set N_e of modern humans between 5,000 and 15,000 (in steps of 1,000) and N_e of Neanderthals between 200 and 2,000 (in steps of 200) after their split at 600,000 years ago. Previous simulation studies used fixed long-term N_e of 10,000 for modern humans and 1,000 for Neanderthals (23, 24) based on whole-genome PSCM estimates (6). Our simulations therefore explore a wider range of possible N_e values centered around the previously used N_e quantities while still keeping the Neanderthal N_e lower than the N_e of modern humans. We deliberately explored a wider range of the Neanderthal N_e because of the higher expected uncertainty in archaic human demographic history compared to modern humans. Furthermore, we investigated the dynamics of introgressed modern human Y chromosomes by simulating gene flow between 1% and 10% (in steps of 1%). Only little is known about the extent of putative gene flow from populations related to early modern humans into Neanderthals but current evidence suggests migration rates on the order of a couple of percent (20, 21); within the range of our simulations.

Given the non-recombining nature of the human Y chromosome, we implemented a simplified model of a genomic structure in which the only parameter of interest is the total amount of sequence under selection (“functional” sequence, ranging from 100 kb to 2 Mb in steps of 100 kb), and set the recombination rate to zero. Furthermore, because the

amount of deleterious variation accumulated on both lineages is directly related to the time they have been separated from each other, we simulated gene flow from modern humans into Neanderthals between 150,000 to 450,000 years ago in steps of 25,000 years (this time range for gene flow encompasses the times inferred by (7, 21, 64) and our own study), assuming a fixed split time of 600,000 years ago. We fixed the deleterious mutation rate at $1.85e-08$ per base pair per generation (4), and sampled the selection coefficients of new mutations from a DFE of non-synonymous changes (65).

Exploring the parameter grid

The large size of the combined parameter space makes exhaustive simulation across the entire parameter grid challenging. Moreover, as we currently lack Y chromosomes of older Neanderthals which would give us information about the dynamics of introgressed modern human Y chromosomes over time, fitting the parameters to the data (represented only by three modern human-like Y chromosomes of Neanderthals around 40 ky old) is currently impossible. Nevertheless, to gain some insights into the probability of replacement given the defined parameter ranges, we explored the parameter space by splitting it into two batches of simulations:

- In the first batch, we evaluated the effect of admixture time and the amount of sequence under selection on the expected decrease of fitness of Neanderthal compared to modern human Y chromosomes (calculated as described below, equation 8) and the probability of Y chromosome replacement. In these simulations,

we kept the N_e of Neanderthals and modern humans fixed at 1,000 and 10,000, respectively, and assumed introgression from modern humans at 5%. For each combination of parameters, we ran 100 independent replicates.

- In the second batch of simulations, we tested the effect of varying the N_e on Neanderthal and modern human lineages and migration rates between 1% and 10%. For these simulations, we fixed the remaining two parameters, the amount of time between the split and introgression and the amount of sequence under selection, at 200 ky and 500 kb, respectively. We chose these two values because they led to a fixation probability of ~50% (at ~2% Neanderthal fitness decrease) in the first batch of simulations (Fig. 3B) and thus represented a useful reference point for exploring the effect of changing the N_e and admixture rate. In this second batch, we simulated 30 replicates.

In both simulation batches, each individual simulation run included an initial burn-in phase of 70,000 generations ($7 \times$ ancestral N_e of 10,000) to let the simulation reach the state of equilibrium.

Scoring the simulated trajectories based on fitness differences

In our previous study, we found evidence for different modes of selection in different classes of functionally important genomic regions, suggesting that the fitness consequences of mutations vary significantly according to the position of their occurrence (63). Realistic modeling of negative selection and introgression would thus require precise information

about the distributions of fitness effects (DFE), dominance and epistasis for coding, non-coding and regulatory regions. Unfortunately, with the exception of DFE of amino-acid-changing *de novo* mutations affecting autosomal genes (65, 66), little is known about fitness consequences of non-coding and regulatory mutations on the Y chromosome. Furthermore, the impact of Y chromosome structural variation in the context of male fertility is highly significant, but still relatively poorly understood in terms of its DFE (26, 67). Similarly, there is substantial uncertainty in many aspects of paternal demographic histories such as variances in reproductive success between individuals in the ancestors of modern humans as well as Neanderthals, and the true N_e dynamics over time, none of which have been yet investigated due to the lack of sufficient Y chromosome data.

To make our results easier to interpret despite the size of the parameter space, we scored each simulation run (i.e., each frequency trajectory of introgressed modern human Y chromosomes in Neanderthals) with the ratio of fitness values of the average Neanderthal Y chromosome and the average modern human Y chromosome generated by the simulation, calculated just prior to the introgression. This way we collapse many potential parameters into a single measure (how much worse are Neanderthal Y chromosomes compared to modern human Y chromosomes in terms of evolutionary fitness) while generalizing our conclusions to other potential factors that we do not model explicitly but which would also influence the fitness difference.

To calculate the fitness of simulated Y chromosomes, we used the fact that mutations in SLiM behave multiplicatively, i.e. each mutation affects any individual's fitness independently of other mutations. Following basic population genetics theory (22),

if we let the fitness of an individual be W and the fitness of each mutation be w_i , we can define W as $W = \prod w_i = \prod(1 - s_i)$, where i runs across all mutations carried by this individual and s_i is the selection coefficient of the i -th deleterious mutation. We can then transform multiplicative interaction into log-additive interaction by

$$W = \prod w_i = e^{\sum \log w_i} = e^{\sum \log(1-s_i)} \approx e^{\sum(-s_i + \frac{s_i^2}{2})} \approx e^{-\sum s_i}, \quad (8)$$

using standard algebraic manipulation and Taylor expansion under the assumption that we are dealing with weakly deleterious mutations whose selection coefficients (s_i) are expected to be small (22).

In practice, we let each simulation replicate run until modern human introgression into Neanderthals, at which point we saved all Neanderthal and modern human Y chromosomes and their mutations to a SLiM population output file. After introgression, we tracked the frequency of modern human Y chromosomes in Neanderthals for 100,000 years, saving the frequency values at regular time intervals to another output file.

For the downstream analyses, we calculated the fitnesses of all Neanderthal and modern human Y chromosomes from the population output files saved in the first step using equation (8) and calculated the ratio of the mean fitness values in both populations – this measure quantifies the decrease in fitness of Neanderthal Y chromosomes compared to modern human Y chromosomes. The distributions of simulated fitness ratios across the two-dimensional parameter grids evaluated for the two simulation batches are shown in Figs. S27 and S30.

To analyze the probability of replacement in Neanderthals over time, we scored the simulated modern human Y chromosome frequency trajectories with the calculated fitness decrease obtained from the simulation. Then, we estimated the expected probability of replacement of the original Neanderthal Y chromosomes over time by counting the proportion of the simulated trajectories (over all replicates trajectories for each combination of parameters) in which the introgressed modern human Y chromosomes reached fixation in each time point. These trajectories of replacement probabilities are shown in Fig. 3B, further averaged by bins of the calculated fitness decrease (Figs. S27 and S28). Fig. S29 shows these probabilities expanded from the single fitness reduction score back on the two-dimensional parameter grid, in a snapshot at 50 ky after introgression. Similarly, Figs. S31 and S32 summarize the probability of replacement in the context of simulations exploring the effect of N_e on both lineages and the rate of modern human introgression. In general, as expected, longer times of separation between Neanderthals and modern humans and larger mutation targets increase the genetic load of Neanderthals (Fig. S27) which, in turn, accelerates their replacement by modern human Y chromosomes (Fig. S28). Similarly, larger differences between the long-term N_e of Neanderthals and modern humans result in lower efficacy of purifying selection in Neanderthals and, consequently, in lower fitness of Neanderthal Y chromosomes (Fig. S30) and acceleration of the replacement dynamics (Figs. S31 and S32).

It is important to note that many different combinations of parameters will lead to similar fitness differences between Neanderthal and modern human Y chromosomes and, consequently, to similar introgression dynamics (Figs. S27 and S28). This is because what ultimately determines the trajectory of introgressed Y chromosomes is how much more fit

they are compared to Neanderthal Y chromosomes (as measured by the fitness difference), regardless of the combination of parameters that lead to this fitness difference.

Finally, we note that although the simulation setup presented here is specific to Y chromosomes, the conclusions based on the abstract measure of fitness reduction could, in principle, be generalized even to the case of mtDNA introgression. This is because our fitness reduction measure applies generally to any non-recombining locus and can be, in fact, interpreted as analogous to a selection coefficient of an allele in a single-locus selection (22).

Supplementary figures

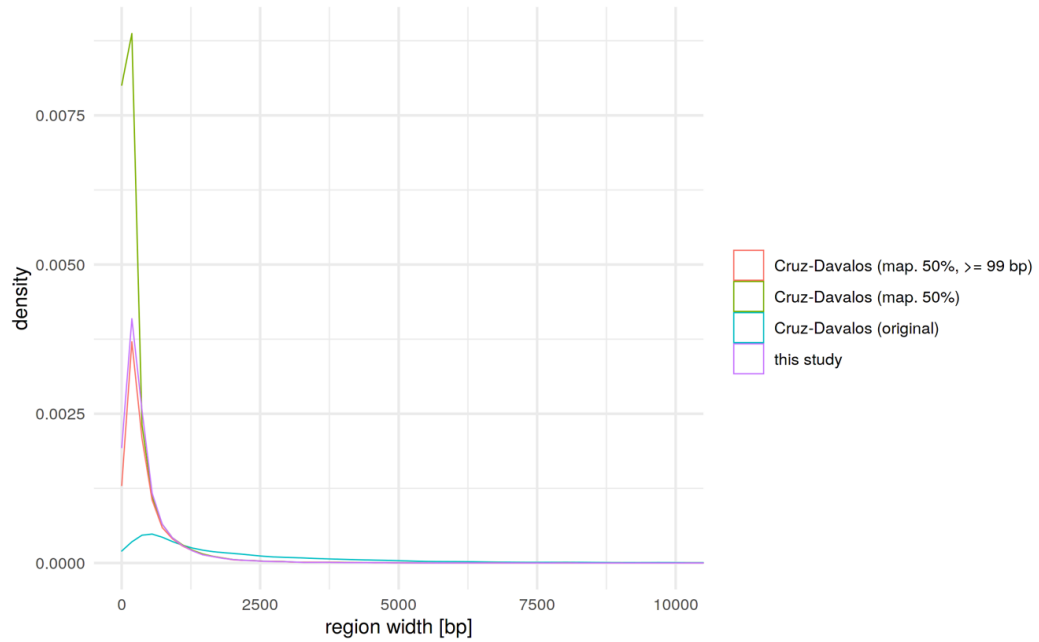


Fig. S1. Distributions of capture region lengths in our study and in Cruz-Dávalos *et al.* (35).

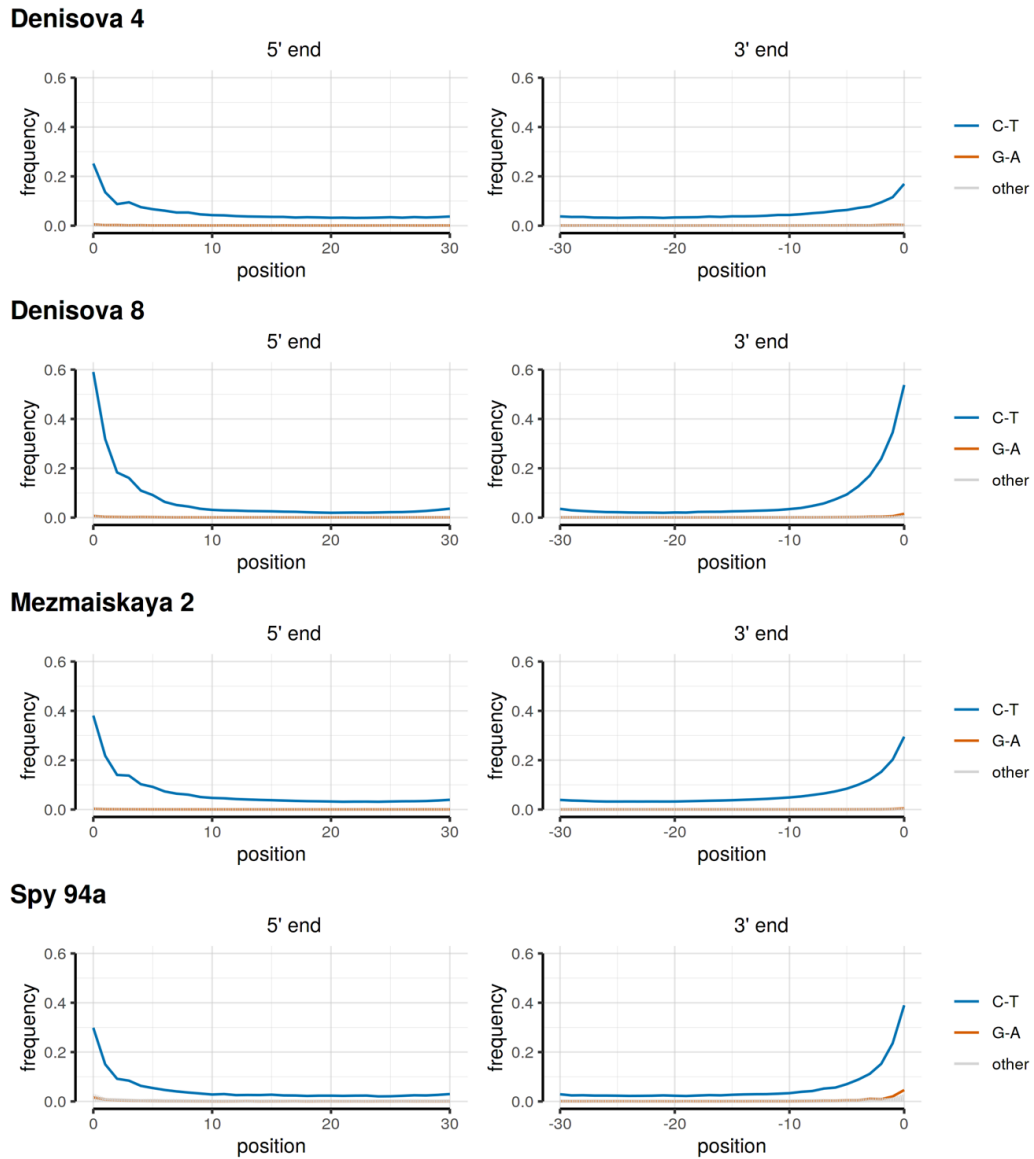


Fig. S2. Patterns of ancient DNA damage in non-UDG-treated sequences captured using the 6.9 Mb capture array.

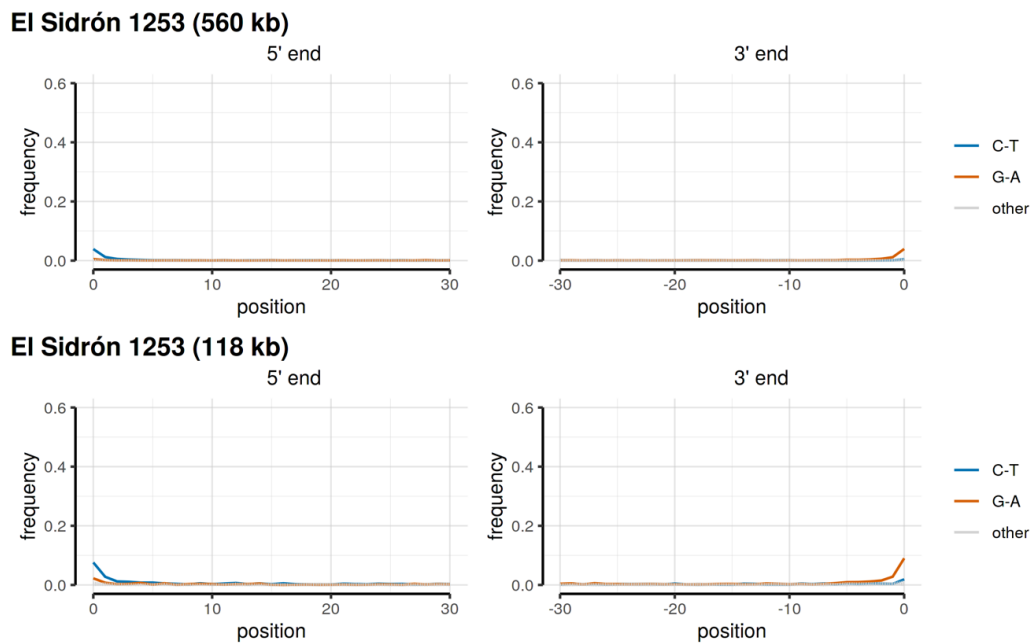


Fig. S3. Patterns of ancient DNA damage in UDG-treated sequences from the *El Sidrón 1253* individual. Top row shows deamination patterns in the 560 kb capture generated for our study, bottom row shows deamination patterns in previously published Y chromosome sequences from the exome capture of the same individual (10, 16).

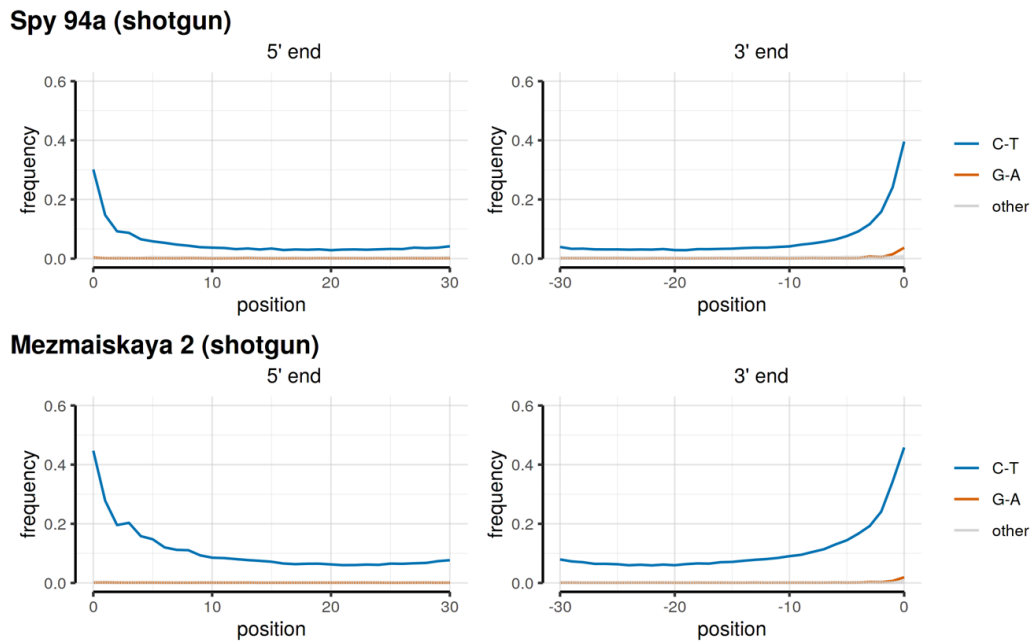


Fig. S4. Patterns of ancient DNA damage in non-UDG-treated shotgun sequences of *Spy 94a* and *Mezmaiskaya 2*. Figures show results from previously published sequence data (13), using sequences within the 6.9 Mb Y chromosome capture target.

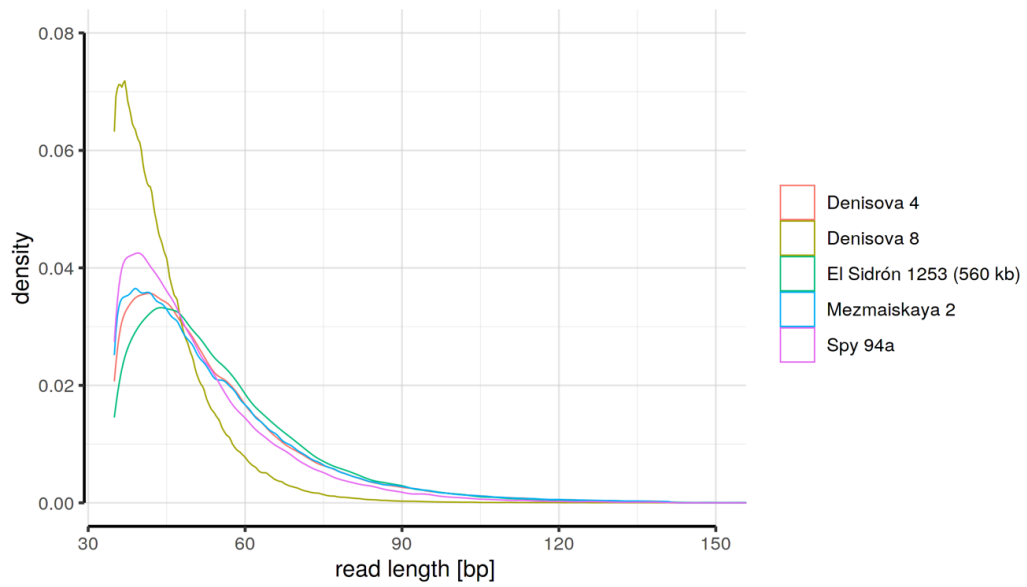


Fig. S5. Distributions of read lengths calculated using sequences within the 6.9 Mb target regions.

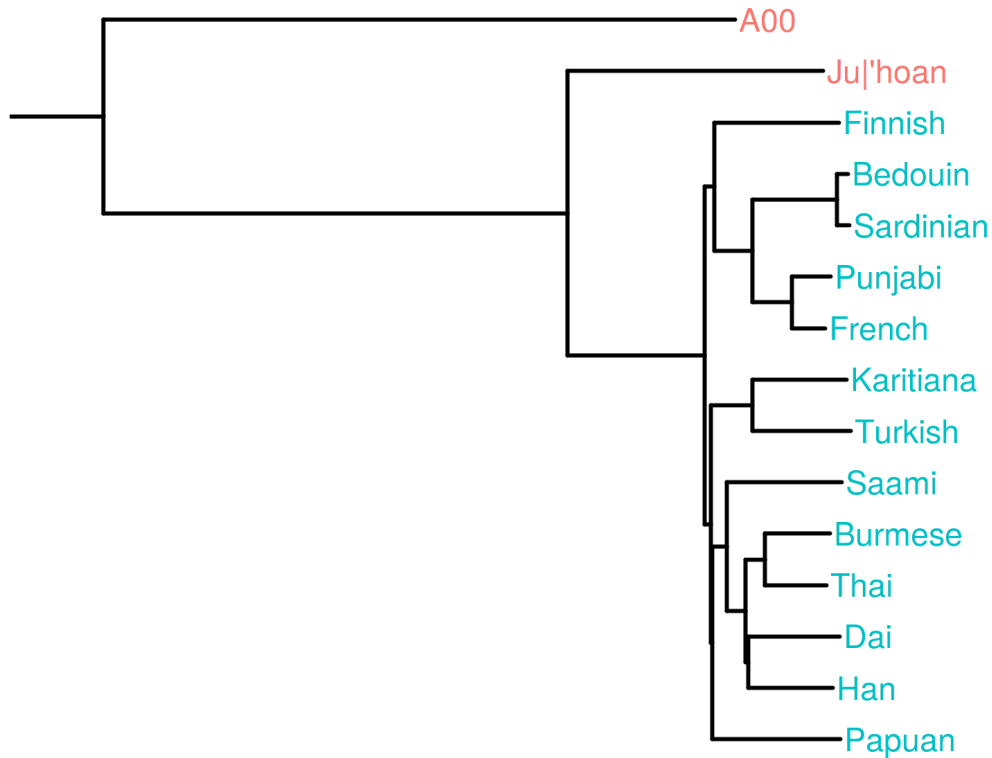


Fig. S6. Phylogenetic tree demonstrating the definition of positions informative about modern human contamination. Red branches represent lineages which we required to carry one state at positions where the blue lineages carry a different state. Therefore, red branches represent the ancestral states and blue branches represent the derived state. The tree was rooted with a chimpanzee Y chromosome as an outgroup (cropped for plotting purposes).

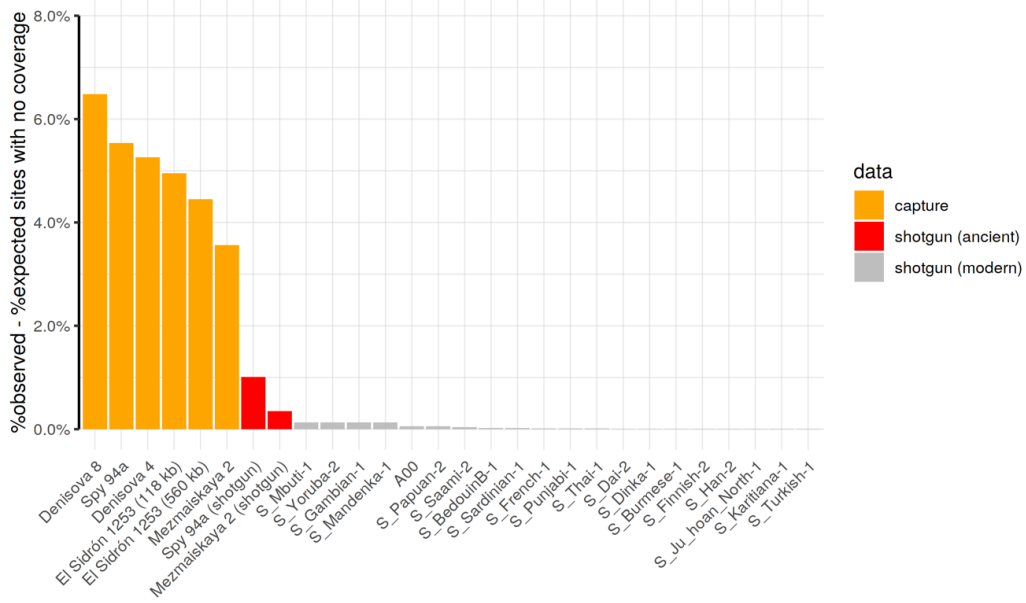


Fig. S7. Differences between expected and observed counts of sites without any sequencing coverage. Exact values are reported in Table S7. Data were filtered by excluding sites with coverage higher than 98% quantile of the coverage distribution.

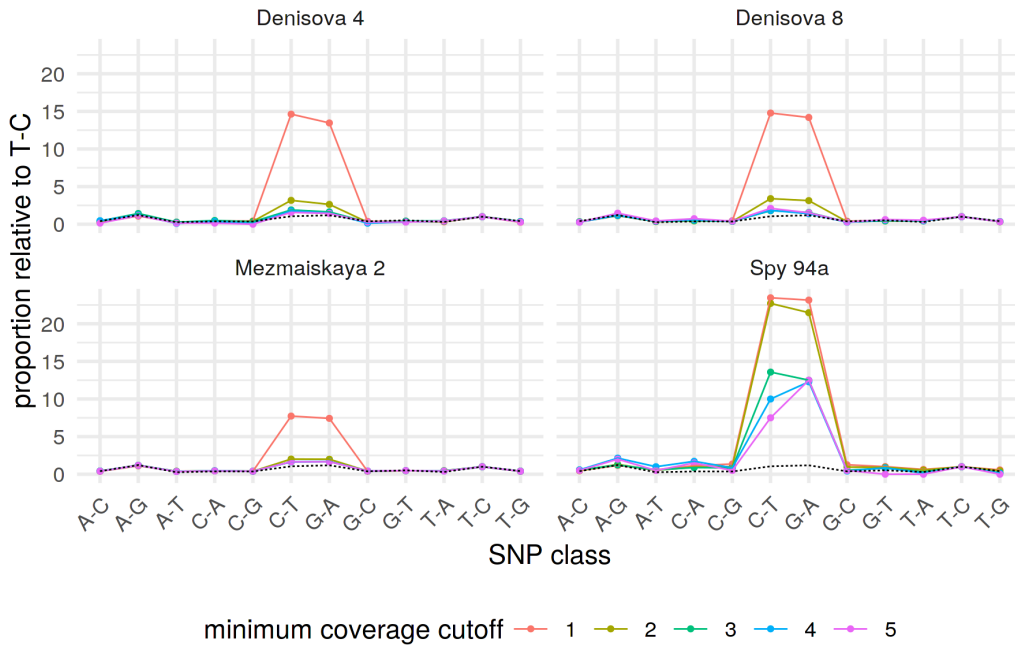


Fig. S8. Frequencies of observed polymorphisms normalized by the observed frequency of T-C polymorphism. SNP classes were counted for each archaic human Y chromosome (one panel each) and all counts were normalized by dividing them with observed counts of T-C SNPs. Dotted line shows an expectation based on SNP proportions observed in Y chromosomes of the following SGDP individuals: *S_French_1*, *S_Papuan_2*, *S_Burmese_1*, *S_Thai_1*, *S_Sardinian_1*.

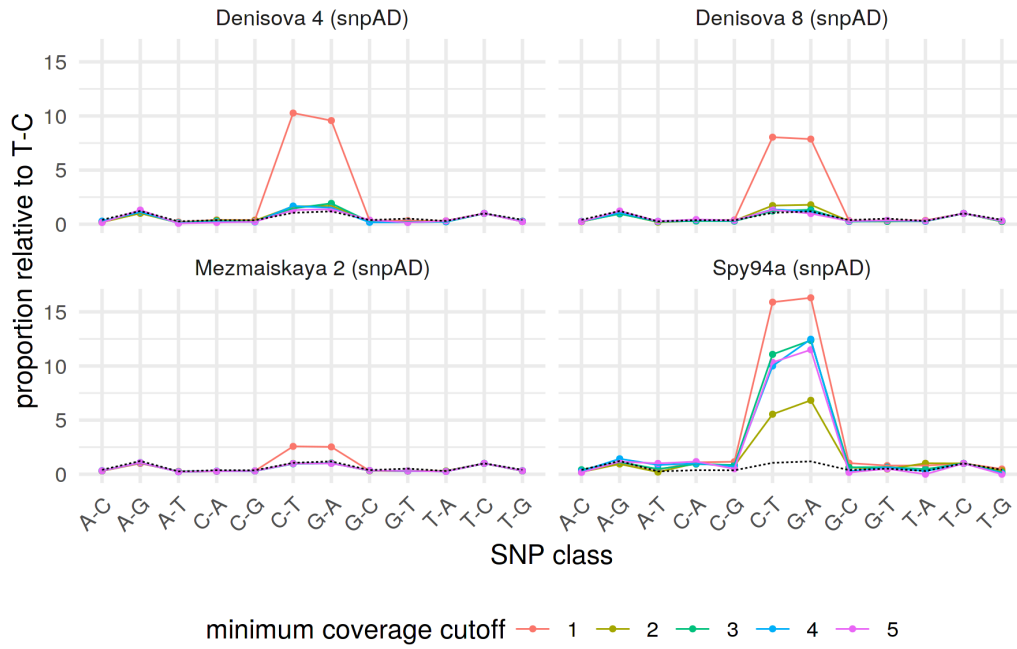


Fig. S9. Frequencies of observed polymorphisms normalized by the observed frequency of T-C polymorphism. SNP classes were counted for each archaic human Y chromosome (one panel each) and all counts were normalized by dividing them with observed counts of T-C SNPs. Dotted line shows an expectation based on SNP proportions observed in Y chromosomes of the following SGDP individuals: *S_French_1*, *S_Papuan_2*, *S_Burmese_1*, *S_Thai_1*, *S_Sardinian_1*.

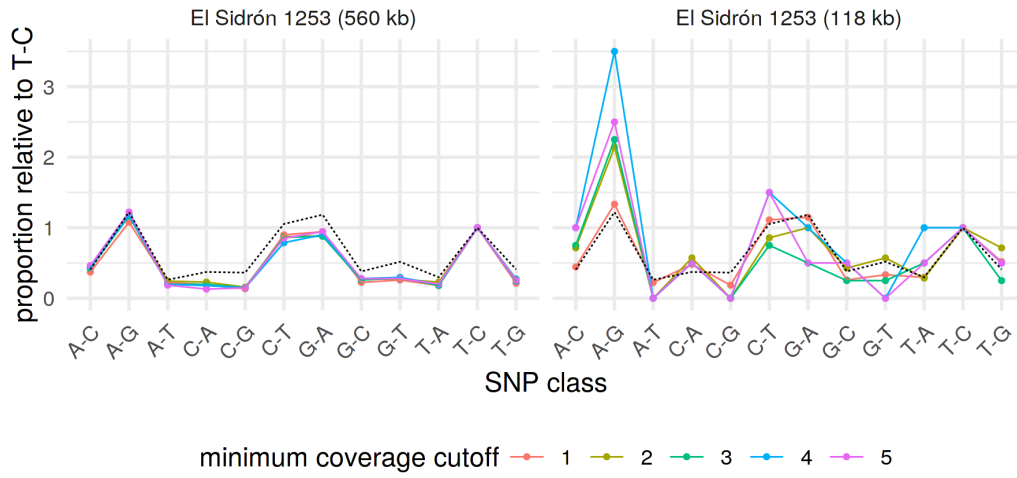


Fig. S10. Frequencies of observed polymorphisms normalized by the frequency of T-C polymorphism. SNP classes were counted for each archaic human Y chromosome (one panel each) and all counts were normalized by dividing them with observed counts of T-C SNPs. Dotted line shows an expectation based on SNP proportions observed in Y chromosomes of the following SGDP individuals: *S_French_1*, *S_Papuan_2*, *S_Burmese_1*, *S_Thai_1*, *S_Sardinian_1*.

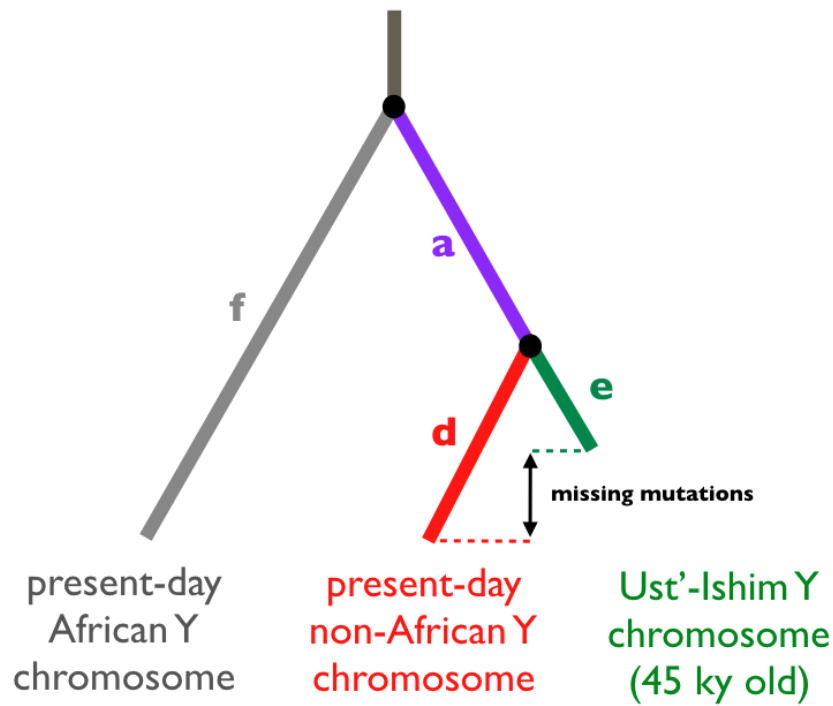


Fig. S11. Schematic of the branch-counting method to estimate the mutation rate and split times of African and non-African Y chromosomes. Accurate knowledge of the age of the *Ust'-Ishim* individual (17) makes it possible to estimate the mutation rate within the 6.9 Mb target capture regions. We use the number of mutations missing on the *Ust'-Ishim* lineage since this individual died (45 kya) and compare it to another non-African Y chromosome, i.e. the quantity $d - e$. We used this mutation rate to calculate the TMRCA between a pair of non-African and African Y chromosomes as the total length of the branches $a + d$.

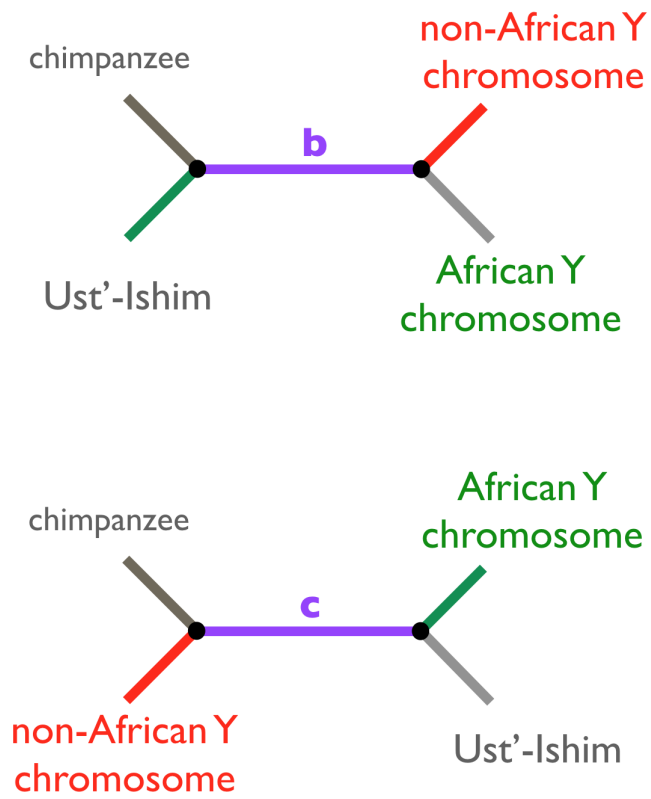


Fig. S12. Two alternative site patterns which are discordant with the true phylogenetic relationship. Given that the true phylogenetic relationship is that shown in Fig. S11, these alternative branch patterns must be a result of double mutations or genotype calling errors.

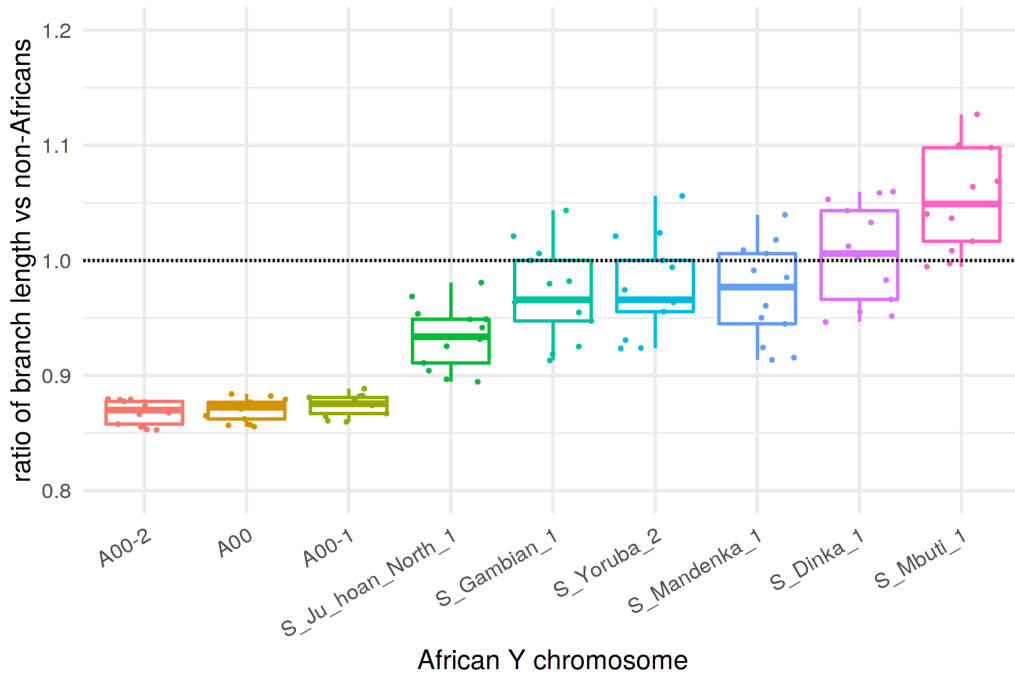


Fig. S13. Branch length differences between African Y chromosomes and a panel of 13 non-African Y chromosomes. Ratios were calculated by creating an alignment of chimpanzee, African and non-African Y chromosomes and taking the ratio of the number of derived alleles observed in an African (x-axis) and the number of derived alleles in each of the individual non-Africans (dots, Table S10). “A00” represents a merge of sequences of two lower coverage Y chromosomes, A00-1 and A00-2 (Table S4).

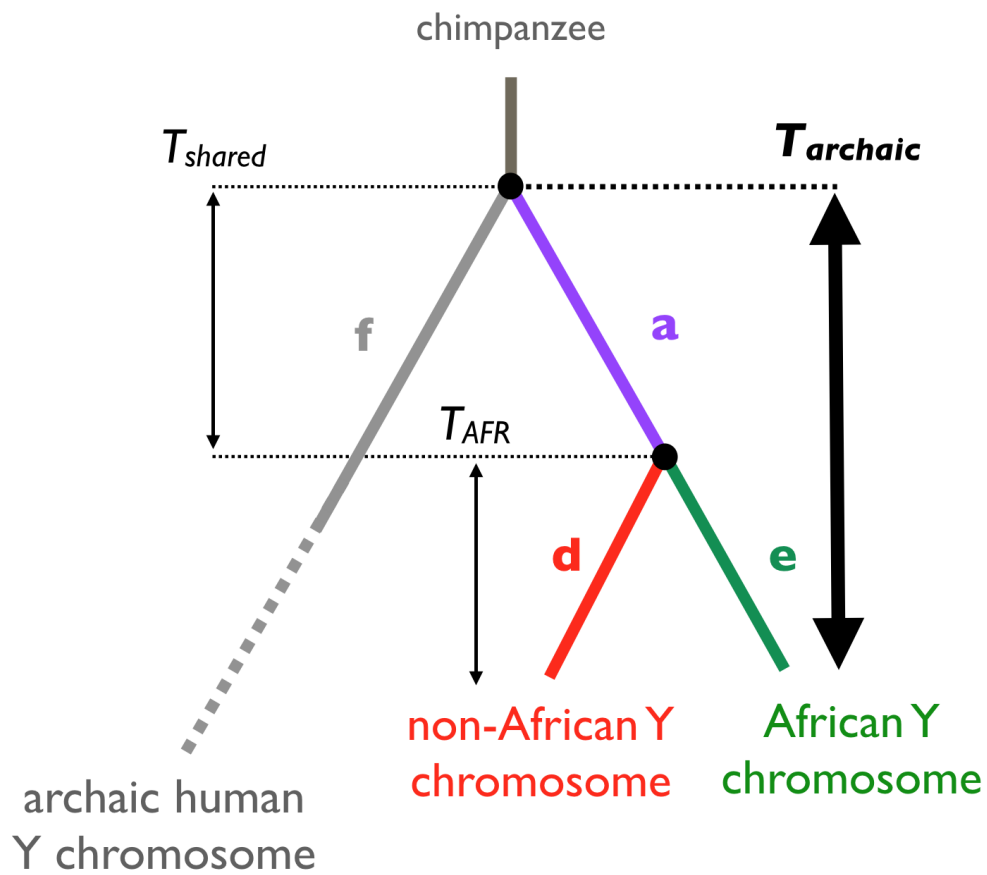


Fig. S14. Branch-counting method to estimate the TMRCA of archaic and present-day human Y chromosomes. As explained in the text, we can decompose the quantity of interest ($T_{archaic}$, thick arrow) using two quantities T_{shared} and T_{AFR} and express it simply as a factor of T_{AFR} , which can be accurately estimated using high-quality modern human Y chromosome sequences.

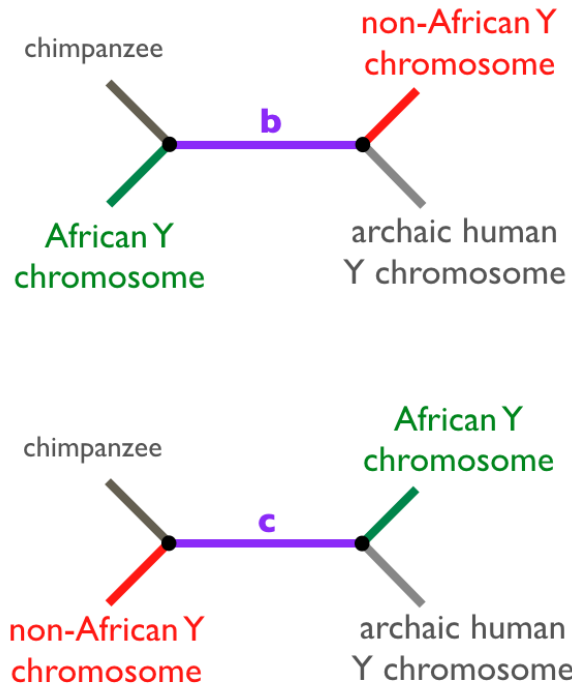


Fig. S15. Alternative tree topologies with two additional possible branch counts *b* and *c*. These topologies are incongruent with the true phylogenetic trees and the branch counts *b* and *c* are most likely a result of back mutations or sequencing errors.

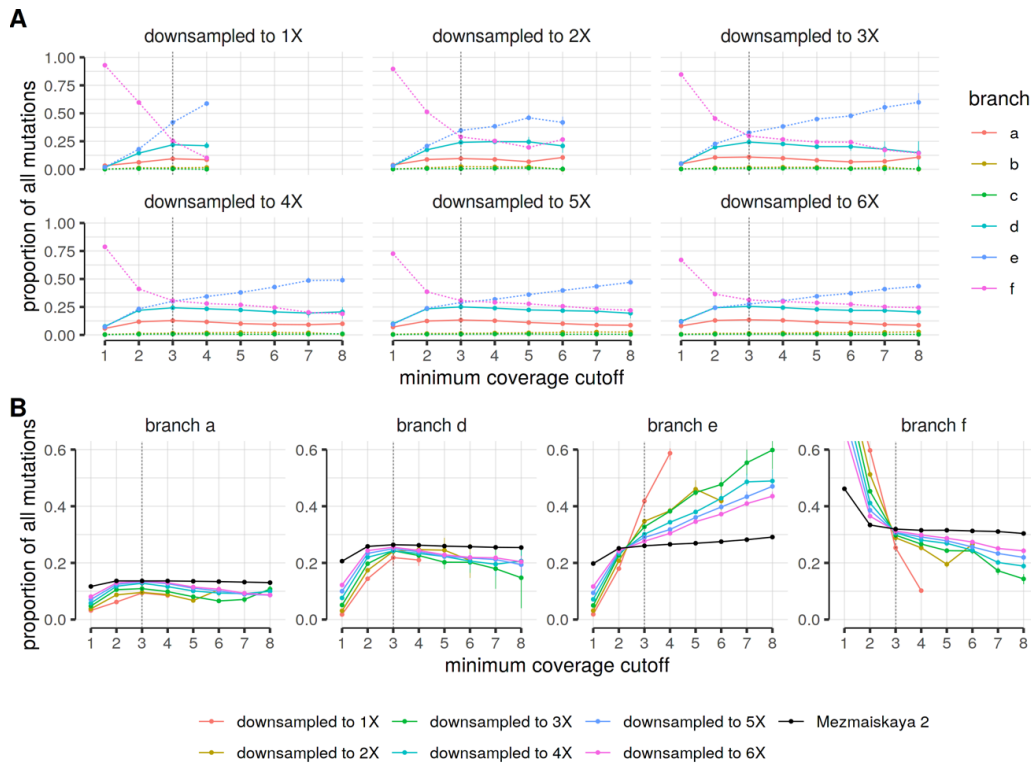


Fig. S16. Relative proportion of branch lengths in downsampled *Mezmaiskaya 2* data as a function of minimum coverage cutoff. (A) Panels show results for 14.3X *Mezmaiskaya 2* downsampled down to 1X, 2X, ... 6X coverage. (B) Same as in panel (A) but partitioned per branch. Black solid lines show expectations based on the full *Mezmaiskaya 2* data. Increased relative proportions of the *f* branch lengths are due to false polymorphisms at low coverage cutoffs. Branch length proportions (labeled as in Fig. S14) were calculated as $\frac{a}{N}, \frac{b}{N}, \dots, \frac{f}{N}$, where $N = a + b + \dots + f$. Vertical dotted lines indicate a 3X lower coverage cutoff used throughout our study. Branch counts were averaged over pairs of 13 non-Africans and the A00 African Y chromosome. Analysis is based on all classes of polymorphisms.

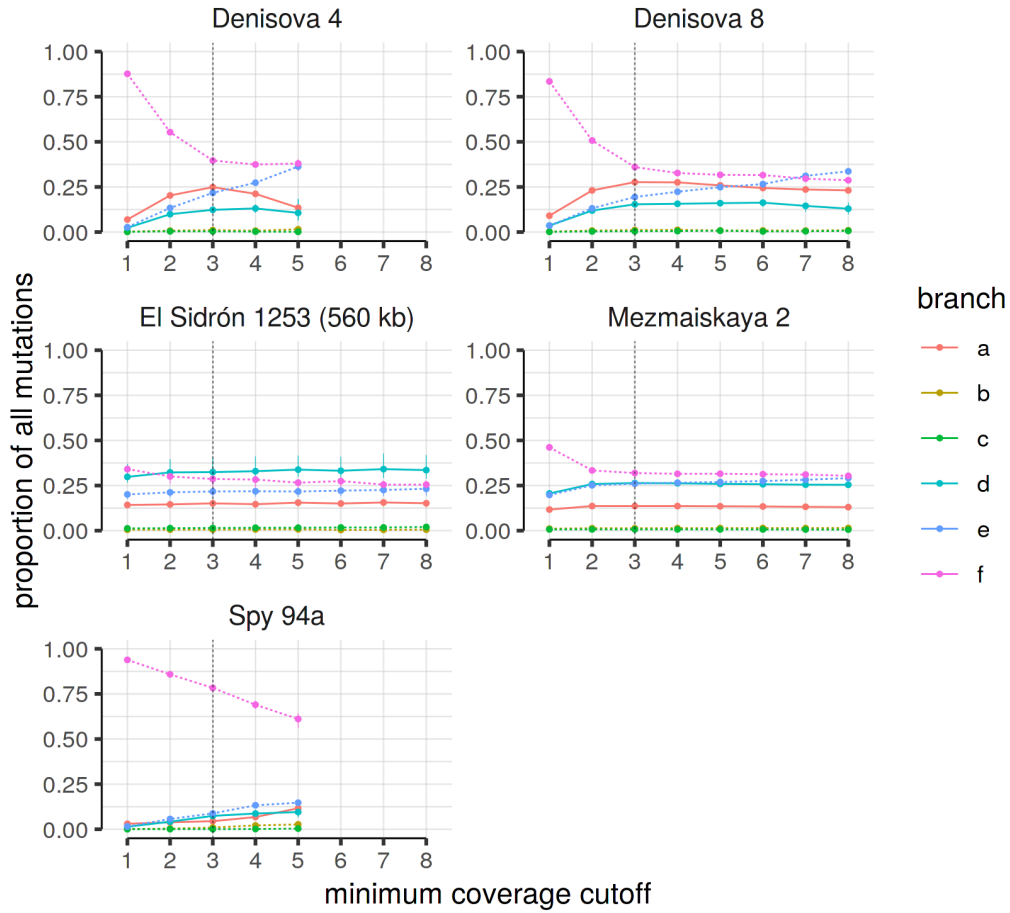


Fig. S17. Relative proportion of branch lengths as a function of minimum coverage support required for each genotype call. Branch length proportions (colored lines) were calculated as $\frac{a}{N}, \frac{b}{N}, \dots, \frac{f}{N}$, where $N = a + b + \dots + f$. Vertical dotted lines indicate a 3X lower coverage cutoff used throughout our study. With the exception of *El Sidrón 1253*, all panels show results for the 6.9 Mb capture data. Analysis is based on all polymorphisms.

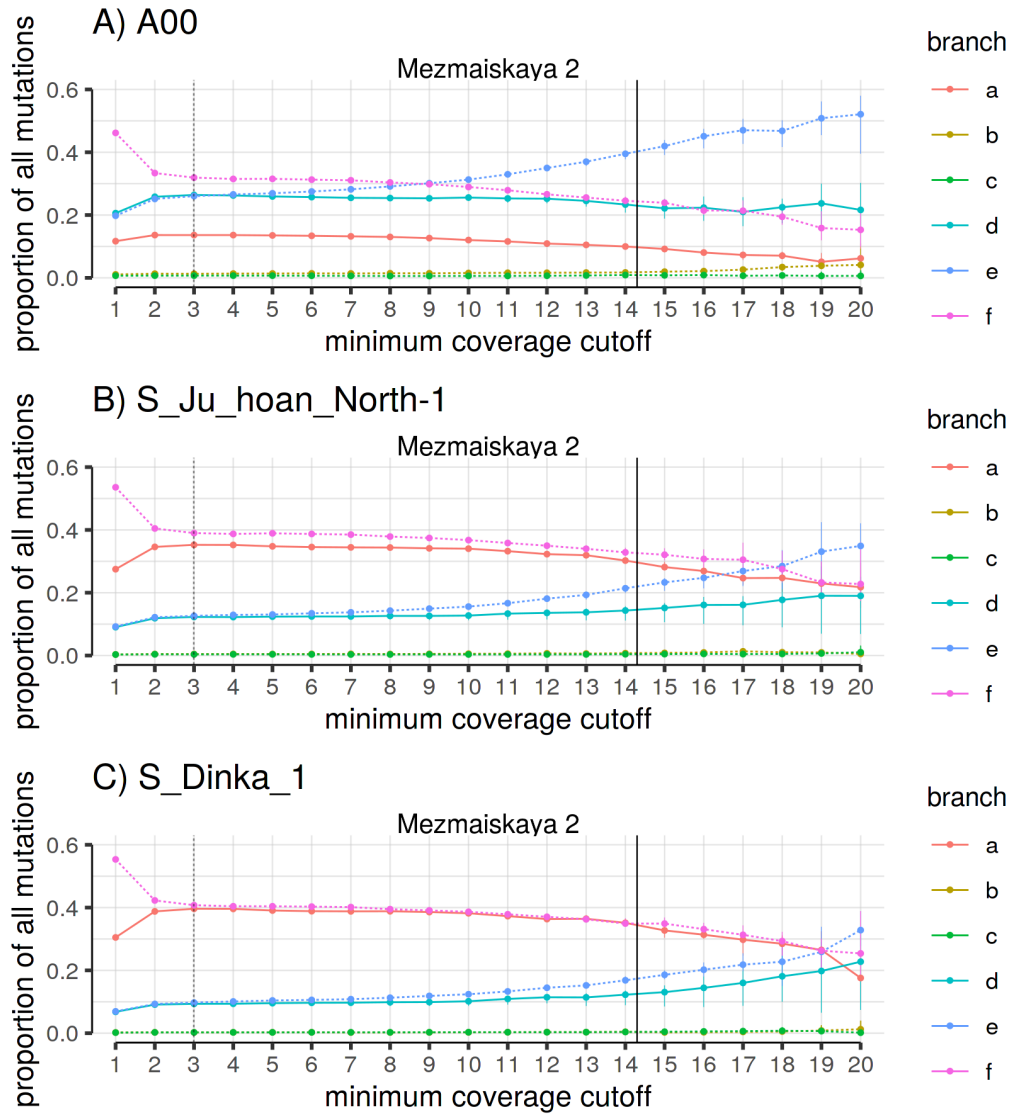


Fig. S18. Relative proportions of branch lengths in the 14.3X *Mezmaiskaya 2* capture data as a function of minimum coverage support required for each genotype call. *Mezmaiskaya 2* was chosen for this example because its high coverage makes the branch proportion patterns stand out more clearly. Panels (A), (B) and (C) show results based on three different African Y chromosomes used to define branch *e* (Fig. S14). Branch length

proportions (colored lines) were calculated as $\frac{a}{N}, \frac{b}{N}, \dots, \frac{f}{N}$, where $N = a + b + \dots + f$, using the complete 6.9 Mb capture data of *Mezmaiskaya 2*. Vertical dotted lines indicate a 3X lower coverage cutoff used throughout our study. Analysis is based on all polymorphisms. Vertical solid lines indicate the average coverage in the *Mezmaiskaya 2* individual (14.3X).

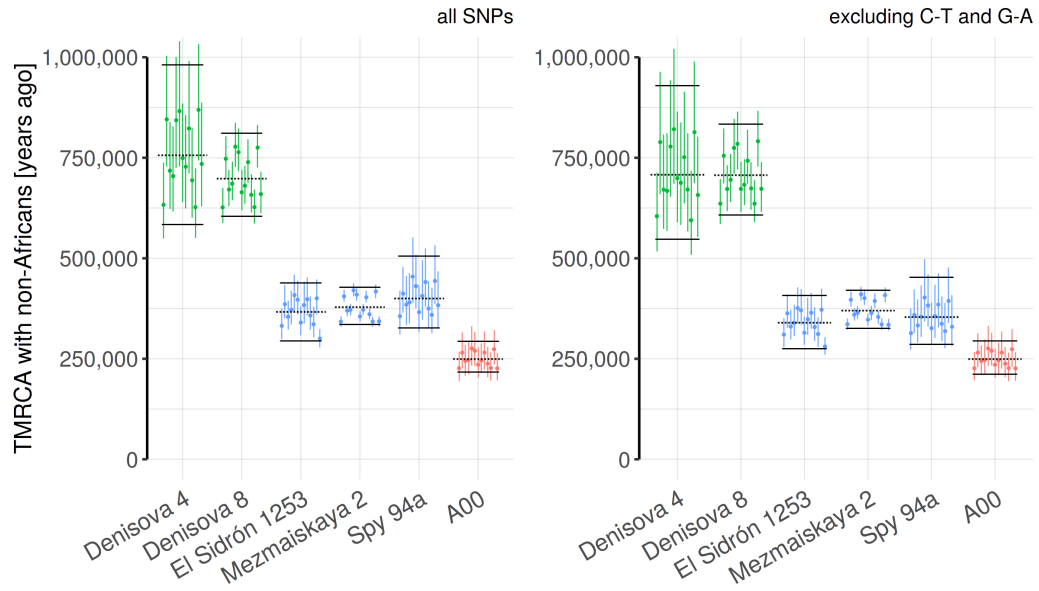
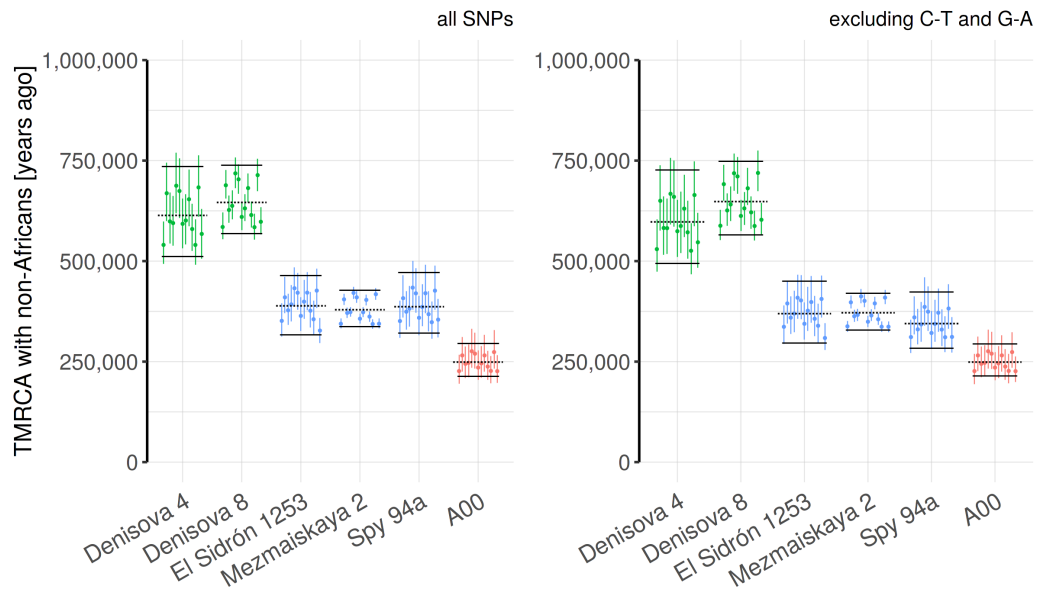
A) New TMRCA statistic**B) TMRCA statistic of Mendez et al.**

Fig. S19. Comparison of TMRCA estimates obtained using the new statistic and the original approach used in the analysis of the *El Sidrón 1253* Neanderthal. (A) Estimates using our new, more robust TMRCA estimate. (B) Original calculation proposed

by *Mendez et al. (10)*. Shown are estimates based on all polymorphisms (left panels) and excluding C-to-T or G-to-A polymorphisms which are more likely to be caused by aDNA damage (right panels).

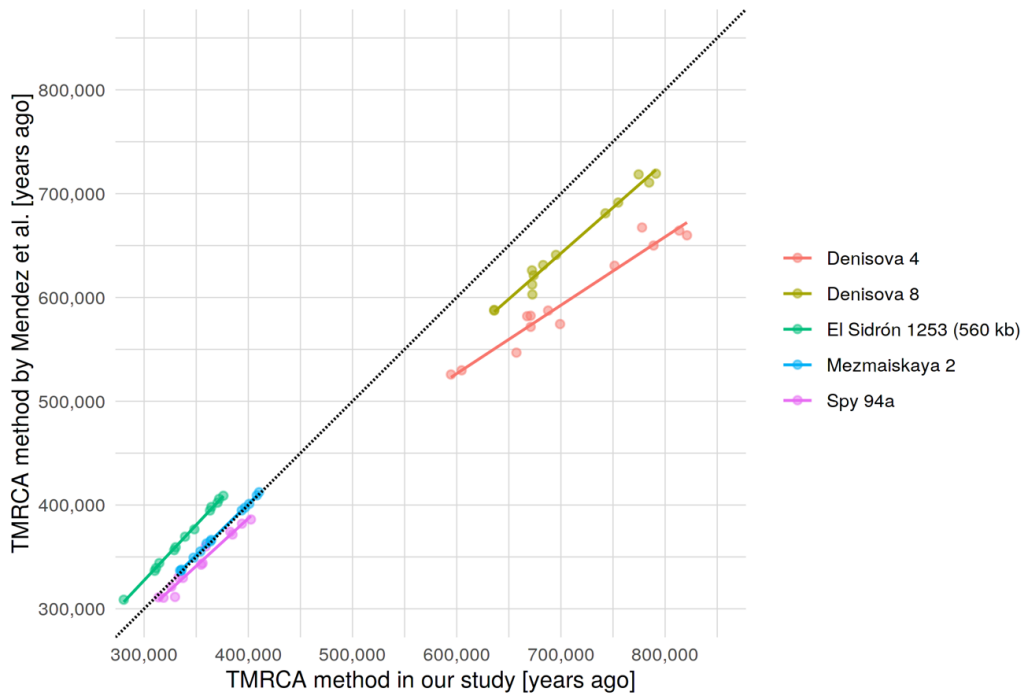
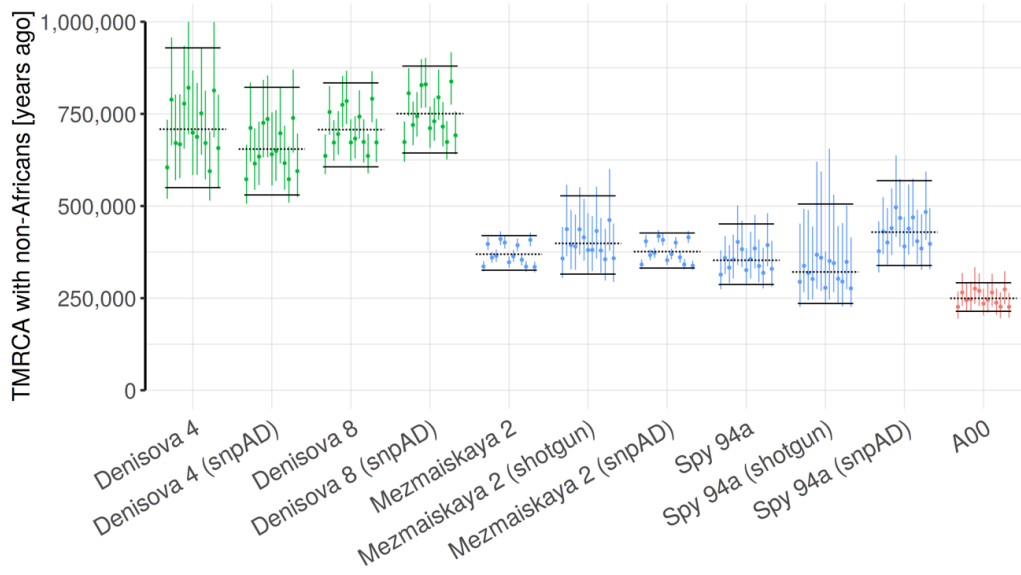
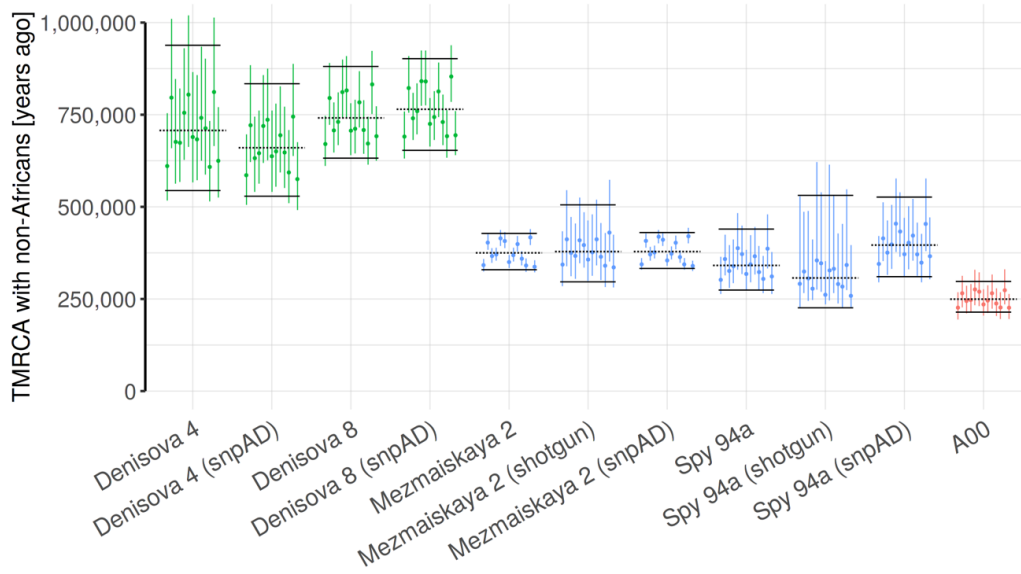


Fig. S20. Comparison of TMRCA estimates obtained using a new statistic and the original approach used in the analysis of the *El Sidrón 1253* Neanderthal. This figure shows the same data as in Fig. S19 (panels on the left) but plotted on the same scatterplot for easier comparison. Dotted black line shows the line of perfect correlation. The TMRCA between the Denisovan individuals and modern humans is slightly underestimated due to a bias in the TMRCA estimation procedure proposed in the original study of the *El Sidrón 1253* Neanderthal (10). Analysis is based on all polymorphisms except C-to-T and G-to-A changes.

A) 50% mappability filter**B) 100% mappability filter****Fig. S21. Detailed evaluation for potential technical biases in our TMRCA estimates.**

Shown are TMRCA results based on different versions of the data (shotgun or capture) or genotype calling methods (snpAD or consensus-based genotype calling method - which is

the default). Panels (A) and (B) show results based on two versions of mappability filters - less strict ("*map35_50%*") and more strict ("*map35_100%*") filters described in the Altai Neanderthal study (6). Analysis is based on all polymorphisms except C-to-T and G-to-A changes.

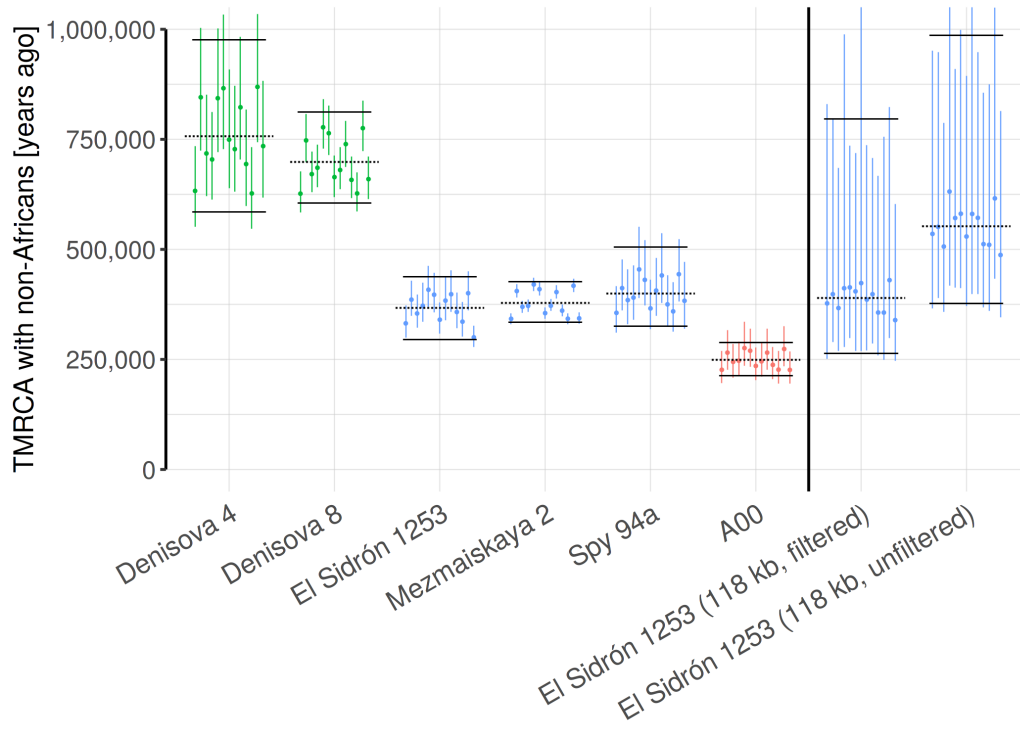


Fig. S22. TMRCA between the *El Sidrón 1253* and modern human Y chromosomes.

TMRCAs obtained for new capture sequence from the *El Sidrón 1253* Neanderthal (~370 kya) differ significantly from the previously published results based on the 118 kb of the coding capture sequence (~600 kya, “118 kb, unfiltered” column on the right from the vertical line). We found that applying stricter filtering criteria results in the same TMRCA values we obtain for the new capture data (“118 kb, filtered” column on the right from the vertical line). Analysis is based on all polymorphisms.

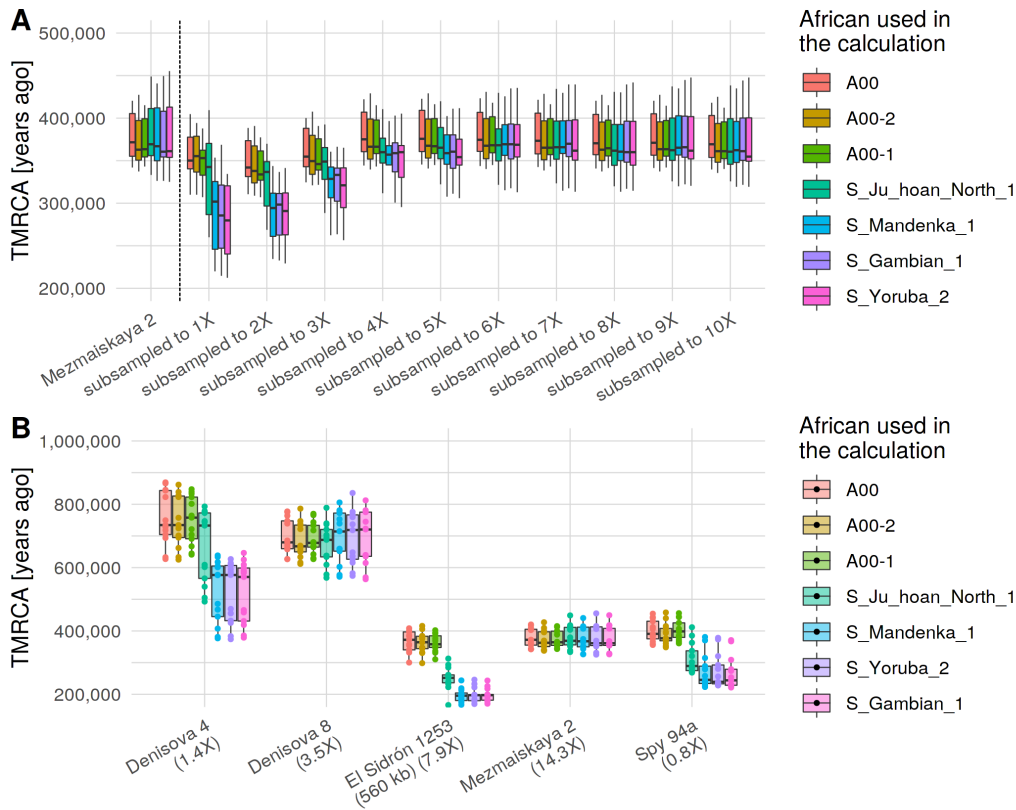


Fig. S23. Estimates of $TMRCA_{archaic}$ for different African Y chromosomes used in the calculation. (A) Results for the high-coverage 14.3X *Mezmaiskaya 2* capture data (left of the vertical line) and its subsets generated by downsampling. A00-based TMRCA estimates are quite stable across the entire range of coverage and match those for the full data. In contrast, estimates based on other, less divergent, African Y chromosomes are heavily biased, and this bias is especially strong for low coverage samples. **(B)** A00-based estimates for *Denisova 4*, *El Sidrón 1253* and *Spy 94a* match those for their higher coverage counterparts (*Denisova 8* and *Mezmaiskaya 2*, respectively) as is required by the topology of the phylogenetic tree (Fig. 2A). However, estimates based on other African individuals show the same bias shown for low coverage samples in panel (A). Dots show TMRCA

estimates based on 13 non-African individuals. Both analyses are based on all polymorphisms.

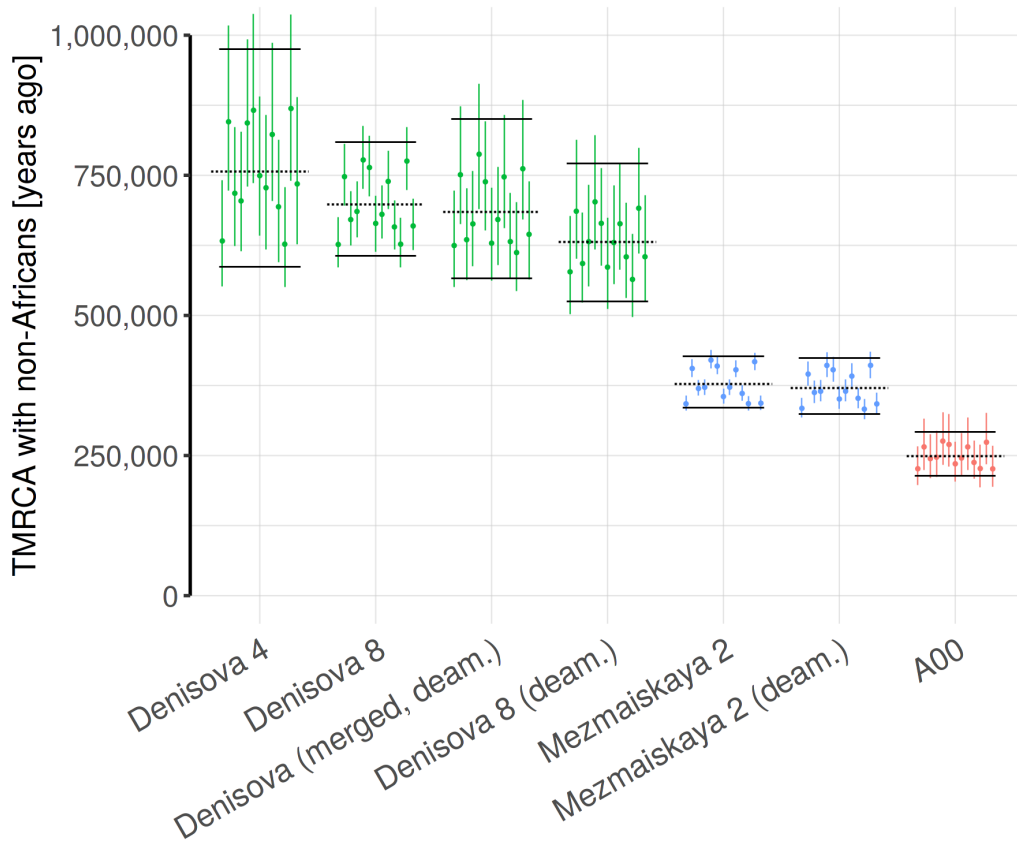


Fig. S24. Comparison of TMRCA estimates using all reads and only reads showing signs of deamination at the three terminal positions (indicated in parentheses). We performed the test only on *Mezmaiskaya 2* and the merged Denisovan data to retain sufficient coverage for the analysis (3.6X coverage for deaminated-only *Mezmaiskaya 2* and 1.5X coverage for deaminated-only merged Denisovan data). Analysis was performed on all SNPs.

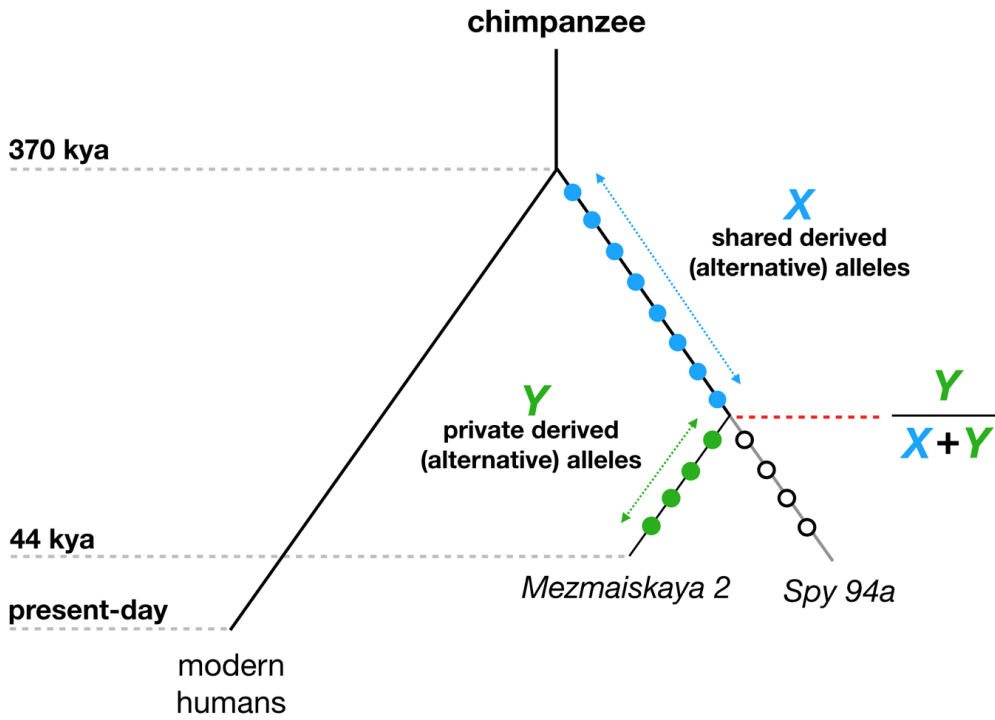


Fig. S25. Estimating the TMRCA of Neanderthal Y chromosomes. Filled circles represent a set of derived (alternative) alleles on the high-coverage *Mezmaiskaya 2* lineage, and are defined as sites at which *Mezmaiskaya 2* Y chromosome carries a different allele than chimpanzee, A00 and French Y chromosomes. Empty circles represent a subset of such sites which show the ancestral (reference) state in *Spy 94a*.

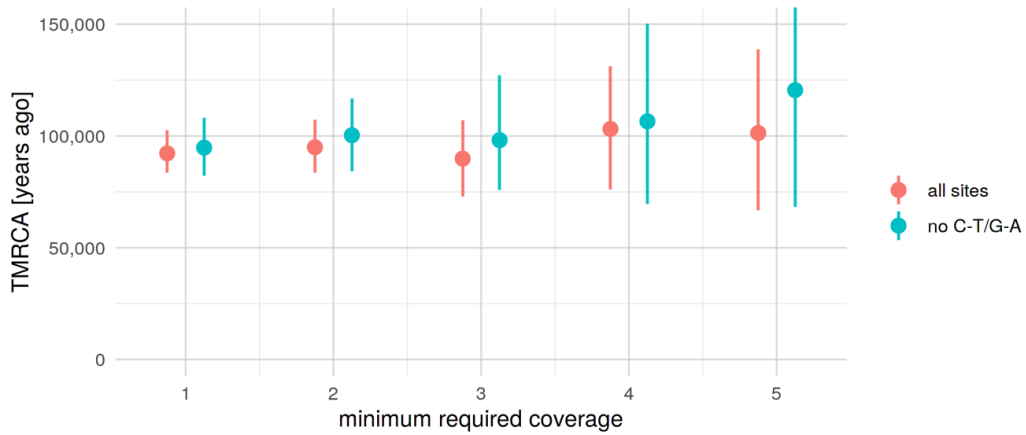


Fig. S26. TMRCA of *Mezmaiskaya 2* and *Spy 94a*. Informative positions (derived alleles in *Mezmaiskaya 2*) were defined using genotype calls in *Mezmaiskaya 2* which passed the standard filtering used throughout our study (minimum coverage of at least three reads, maximum coverage less than 98% quantile of the total coverage distribution). We called the genotypes in the *Spy 94a* Y chromosome at these positions and calculated the TMRCA using the equation (7). We tested the robustness of the estimate to genotype calling errors using different minimum coverage filters for *Spy 94a* (x-axis) and two sets of polymorphisms (colors).

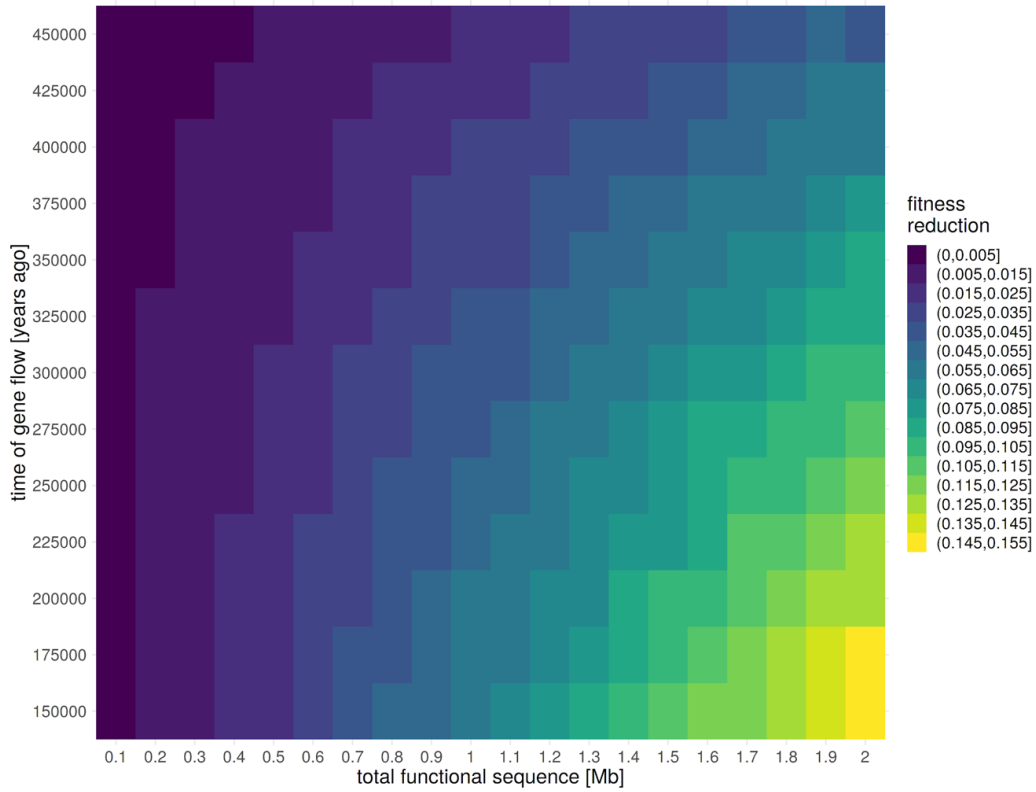


Fig. S27. Decrease in fitness of an average Neanderthal Y chromosome compared to an average modern human Y chromosome. Fitness decrease values were averaged over 100 independent simulation replicates on a grid of two parameters as the ratios of the mean fitness of Neanderthal Y chromosomes to the mean fitness of modern human Y chromosomes (calculated using equation 8). Lighter colors represent lower fitness of Neanderthal Y chromosomes compared to modern human Y chromosomes.

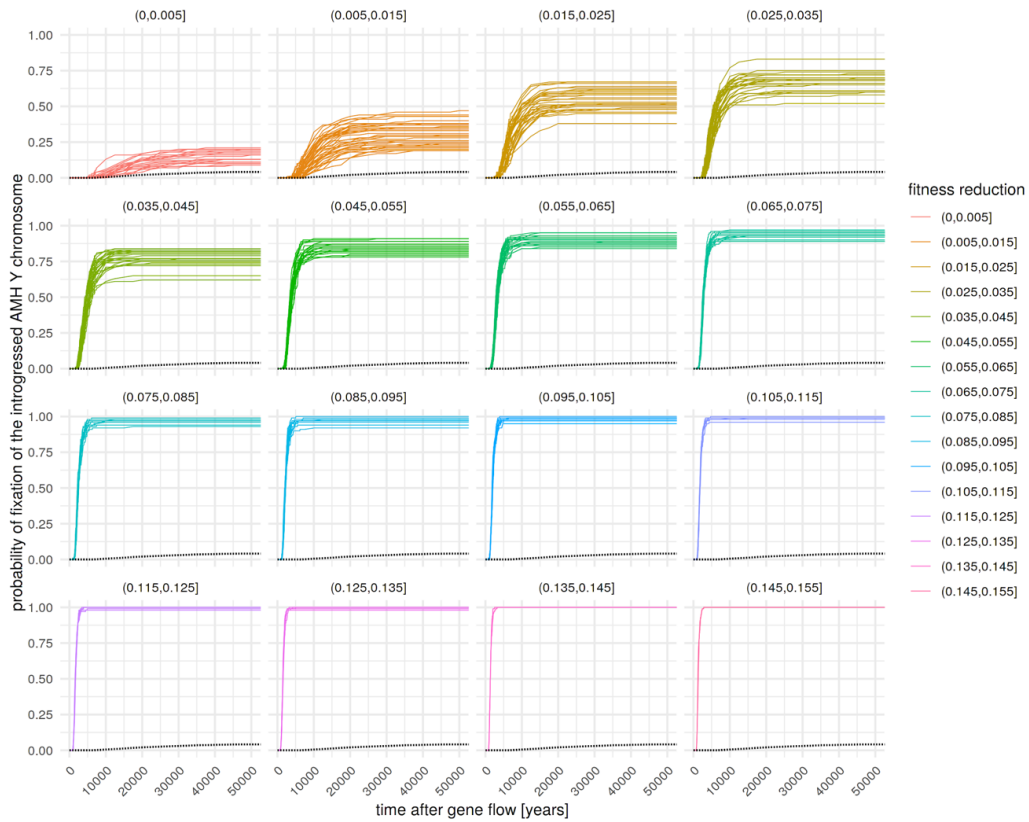


Fig. S28. Trajectories of the probability of fixation of introgressed modern human Y chromosomes in Neanderthals, partitioned by the fitness decrease of Neanderthal Y chromosomes compared to modern human Y chromosomes. Black dotted lines show the probability of replacement under neutrality. All simulations assumed 5% initial gene flow from modern humans into Neanderthals. Fig. 3B was generated from this data after calculating average probabilities across all trajectories in each fitness reduction bin.

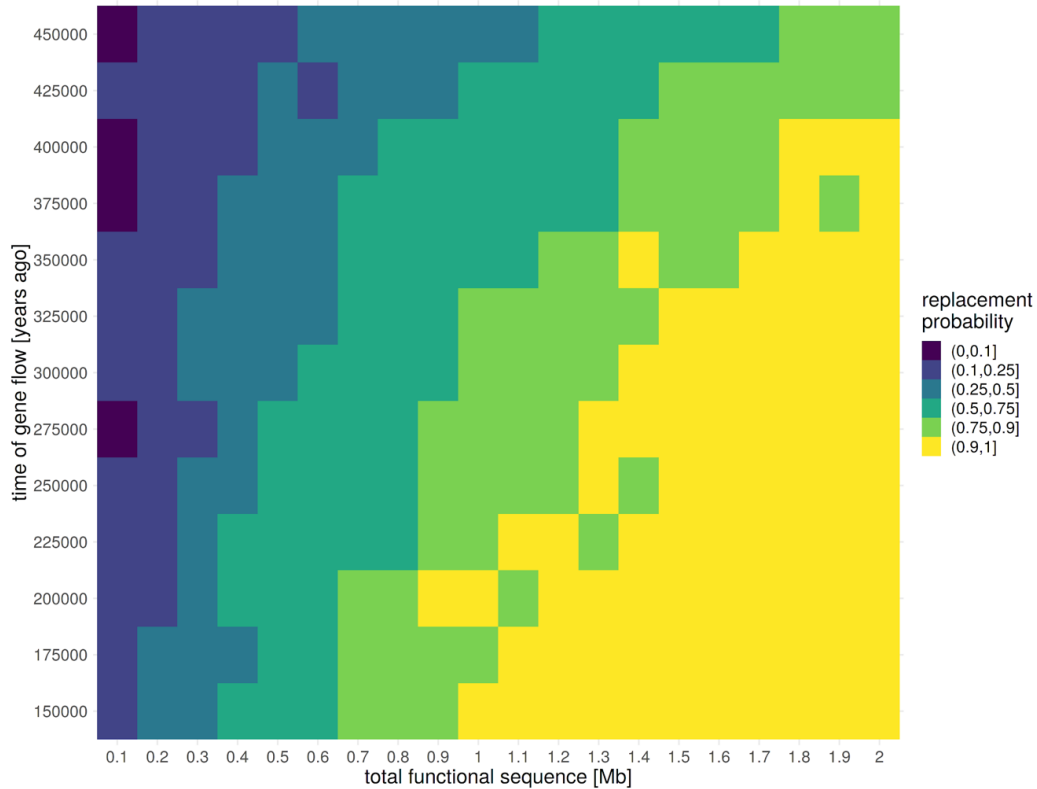


Fig. S29. Probability of replacement of the Neanderthal Y chromosomes at 50 thousand years after gene flow from modern humans. Probabilities represent the proportion of introgressed modern human Y chromosome trajectories that reached fixation in the Neanderthals after 50 thousand years after gene flow out of the total 100 simulation replicates performed for each combination of two-dimensional parameters (Figs. S27 and S28).

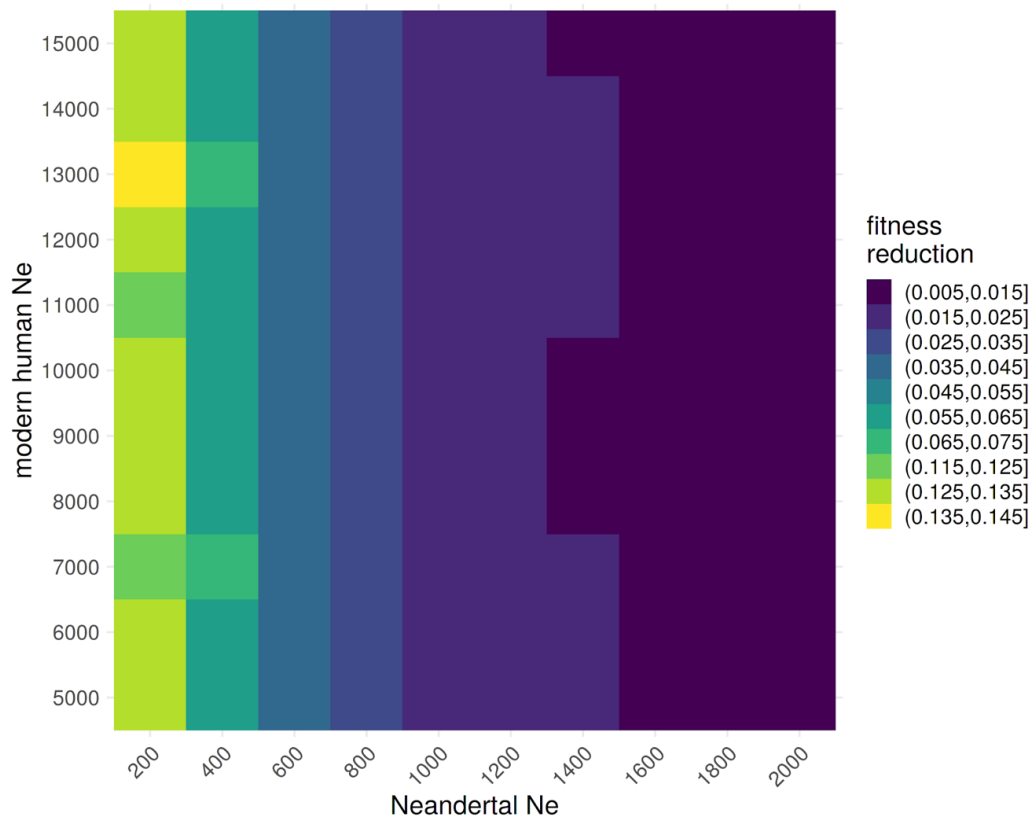


Fig. S30. Decrease in fitness of an average Neanderthal Y chromosome compared to an average modern human Y chromosome. Fitness decrease values were averaged over 30 independent simulation replicates on a grid of two parameters as the ratios of the mean fitness of Neanderthal Y chromosomes to the mean fitness of modern human Y chromosomes (calculated using equation 8). Lighter colors represent lower fitness of Neanderthal Y chromosomes compared to modern human Y chromosomes.

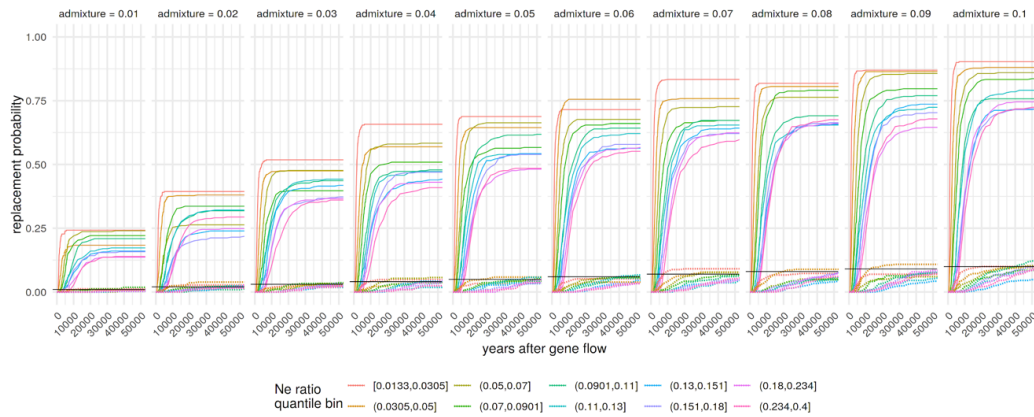


Fig. S31. Trajectories of the probability of replacement in the simulation scenario which varied N_e of Neanderthals and modern humans as well as the rate of introgression. Individual frequency trajectories were first partitioned into equal-sized bins based on the ratio of N_e values of both populations (modern human N_e between 5,000 and 15,000 in steps of 1,000, Neanderthal N_e between 200 and 2,000 in steps of 200) and the trajectory of the probability of replacement was calculated as the proportion of trajectories in each bin which reached the frequency 1.0 (i.e., fixation) in each time point. Dotted lines show expectations of replacement probabilities under neutrality.

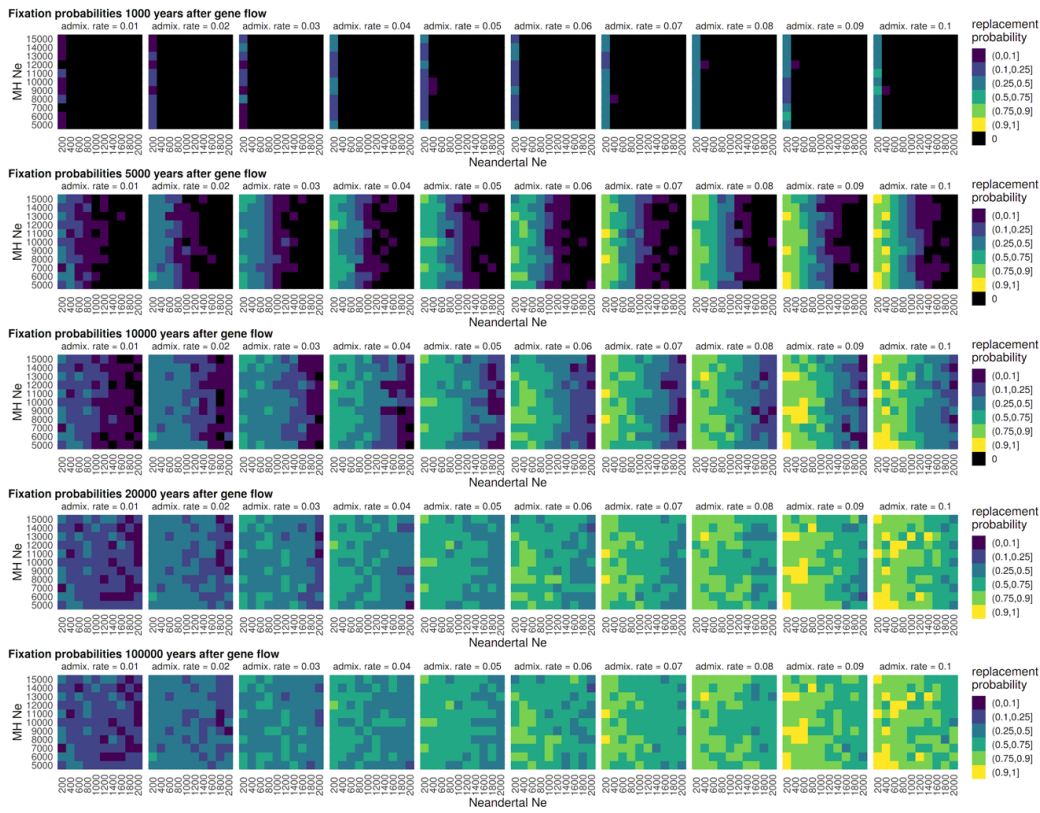


Fig. S32. Probability of replacement of the Neanderthal Y chromosomes at different time points after gene flow from modern humans. Probabilities represent the proportion of introgressed modern human Y chromosomal trajectories that reached fixation in the Neanderthals at a given time point, out of the total 30 simulation replicates performed for each combination of the two N_e parameters.

Supplementary tables

target	total [bp]	# of regions	min [bp]	median [bp]	mean [bp]	max [bp]
entire mappable Y	6,912,728	15,903	99.0	240.0	434.7	9,425.0
~560 kb capture (*)	573,657	1,251	60.0	151.0	458.6	3899.0
~560 kb capture	556,259	1,779	1.0	119.0	312.7	2,829.0
exome subset	118,643	2,519	1.0	3.0	47.1	1,257.0

Table S1. Characteristics of the three sets of Y chromosome capture targets analyzed in our study. “*Exome subset*” refers to a Y chromosome subset of the exome capture sequence generated by Castellano *et al.* and analyzed by Mendez *et al.* (called “filter 1”) (10, 16), “~560 kb capture” refers to target regions originally designed for studying present-day human Y chromosome variation (3), star (*) signifies statistics before intersecting the original set of target regions with the “*map35_50%*” filter (6), “*entire mappable Y*” represents capture regions targeting the entire mappable portion of the human Y chromosome designed for our study.

individual	mean coverage	target
Spy 94a	0.8	6.9 Mb
Denisova 4	1.4	6.9 Mb
Denisova 8	3.5	6.9 Mb
El Sidrón 1253	7.9	560 kb
Mezmaiskaya 2	14.3	6.9 Mb

Table S2. Mean coverage of archaic human Y chromosomes sequenced in this study.

Sites with coverage higher than 98% quantile of the entire distribution were excluded from the calculation.

name	mean coverage	target	study
Spy 94a (shotgun)	0.5	6.9 Mb	(13)
Mezmaiskaya 2 (shotgun)	0.8	6.9 Mb	(13)
El Sidrón 1253 (capture)	3.2	118 kb	(10, 16)

Table S3. Mean coverage of previously published archaic human Y chromosome sequences. The coverage reported for *Spy 94a* and *Mezmaiskaya 2* shotgun sequence data is that of sequences overlapping the 6.9 Mb Y capture regions. The *El Sidrón 1253* libraries were captured using an exome capture array (10, 16) and the coverage reported here is for the ~118 kb exome target capture regions. For each individual, sites with coverage higher than 98% quantile of the entire distribution were excluded from the calculation.

name	mean coverage	target	study
A00-1	8.8	6.9 Mb	(4)
A00-2	12.0	6.9 Mb	(4)
S_Mandenka-1	16.3	6.9 Mb	(18)
S_Yoruba-2	17.0	6.9 Mb	(18)
S_Finnish-2	17.0	6.9 Mb	(18)
S_Punjabi-1	17.1	6.9 Mb	(18)
S_Sardinian-1	18.1	6.9 Mb	(18)
S_Dai-2	19.8	6.9 Mb	(18)
S_Gambian-1	20.1	6.9 Mb	(18)
Ust'-Ishim	20.1	6.9 Mb	(17)
S_Mbuti-1	20.3	6.9 Mb	(18)
S_Dinka-1	20.8	6.9 Mb	(18)
S_Han-2	20.8	6.9 Mb	(18)
A00	20.9	6.9 Mb	(4)
S_BedouinB-1	21.7	6.9 Mb	(18)
S_French-1	21.9	6.9 Mb	(18)
S_Karitiana-1	22.2	6.9 Mb	(18)
S_Turkish-1	22.5	6.9 Mb	(18)
S_Saami-2	22.6	6.9 Mb	(18)
S_Ju_hoan_North-1	22.7	6.9 Mb	(18)
S_Papuan-2	23.2	6.9 Mb	(18)
S_Thai-1	25.1	6.9 Mb	(18)
S_Burmese-1	29.2	6.9 Mb	(18)

Table S4. Mean coverage of modern human Y chromosomes in capture target regions.

Coverage is reported using sequences within the 6.9 Mb target capture regions. For each individual, sites with coverage higher than 98% quantile of the entire distribution were excluded from the calculation.

individual	mean [bp]	median [bp]
Denisova 8	44.8	42
Spy 94a	51.2	47
Mezmaiskaya 2	54.1	49
Denisova 4	54.1	50
El Sidrón 1253 (560 kb)	55.3	51

Table S5. Mean and median values of read length distributions in Fig. S5.

individual	ancestral count	derived count	total	derived/total
Spy 94a	15	1	16	0.0625
Mezmaiskaya 2	189	0	189	0.0000
Denisova 4	14	0	14	0.0000
Denisova 8	90	0	90	0.0000
El Sidrón 1253 (560 kb)	29	0	29	0.0000

Table S6. Counts and proportions of potential ‘contaminant-derived’ non-African alleles in all archaic human Y chromosomes.

name	coverage	observed	expected	difference	data
Denisova 8	3.4820355	0.09555	0.03074	0.06481	capture
Spy 94a	0.8253622	0.49341	0.43808	0.05533	capture
Denisova 4	1.3717806	0.30622	0.25365	0.05257	capture
El Sidrón 1253 (118 kb)	3.2121215	0.08978	0.04027	0.04951	capture
El Sidrón 1253 (560 kb)	7.9165032	0.04479	0.00036	0.04443	capture
Mezmaiskaya 2	14.3494002	0.03561	0.00000	0.03561	capture
Spy 94a (shotgun)	0.5052873	0.61337	0.60333	0.01004	shotgun (ancient)
Mezmaiskaya 2 (shotgun)	0.8248822	0.44174	0.43829	0.00345	shotgun (ancient)
S_Yoruba-2	16.9616115	0.00129	0.00000	0.00129	shotgun (modern)
S_Mbuti-1	20.3360989	0.00129	0.00000	0.00129	shotgun (modern)
S_Mandenka-1	16.3312730	0.00127	0.00000	0.00127	shotgun (modern)
S_Gambian-1	20.1210272	0.00127	0.00000	0.00127	shotgun (modern)
A00	20.8671570	0.00050	0.00000	0.00050	shotgun (modern)
S_Papuan-2	23.1529894	0.00048	0.00000	0.00048	shotgun (modern)
S_Saami-2	22.6298230	0.00033	0.00000	0.00033	shotgun (modern)
S_BedouinB-1	21.7140764	0.00016	0.00000	0.00016	shotgun (modern)
S_Sardinian-1	18.1189362	0.00015	0.00000	0.00015	shotgun (modern)
S_Punjabi-1	17.1469680	0.00013	0.00000	0.00013	shotgun (modern)
S_French-1	21.9031161	0.00013	0.00000	0.00013	shotgun (modern)
S_Thai-1	25.1203313	0.00010	0.00000	0.00010	shotgun (modern)
S_Dai-2	19.8120096	0.00004	0.00000	0.00004	shotgun (modern)
S_Dinka-1	20.7590166	0.00004	0.00000	0.00004	shotgun (modern)
S_Finnish-2	17.0155033	0.00003	0.00000	0.00003	shotgun (modern)
S_Han-2	20.7680004	0.00003	0.00000	0.00003	shotgun (modern)
S_Karitiana-1	22.1963029	0.00003	0.00000	0.00003	shotgun (modern)
S_Turkish-1	22.4638799	0.00003	0.00000	0.00003	shotgun (modern)
S_Ju_hoan_North-1	22.7025717	0.00003	0.00000	0.00003	shotgun (modern)
S_Burmese-1	29.2087547	0.00003	0.00000	0.00003	shotgun (modern)

Table S7. Proportions of expected and observed sites without any coverage.

filter	name	all SNPs		excluding C-to-T/G-to-A	
		count	proportion	count	proportion
unfiltered	Spy 94a (shotgun)	2,664,787	38.5%	2,636,405	38.1%
unfiltered	Spy 94a	3,502,380	50.7%	3,470,671	50.2%
unfiltered	Spy 94a (snpAD)	3,556,151	51.4%	3,522,222	51.0%
unfiltered	Mezmaiskaya 2 (shotgun)	3,822,106	55.3%	3,762,768	54.4%
unfiltered	Denisova 4	4,731,302	68.4%	4,705,860	68.1%
unfiltered	Denisova 4 (snpAD)	4,823,855	69.8%	4,797,798	69.4%
unfiltered	Denisova 8	5,851,332	84.6%	5,828,356	84.3%
unfiltered	Denisova 8 (snpAD)	6,259,573	90.6%	6,235,802	90.2%
unfiltered	Mezmaiskaya 2	6,348,948	91.8%	6,346,684	91.8%
unfiltered	Mezmaiskaya 2 (snpAD)	6,669,431	96.5%	6,667,075	96.4%
min DP \geq 3	Spy 94a (shotgun)	150,642	2.2%	150,606	2.2%
min DP \geq 3	Mezmaiskaya 2 (shotgun)	450,609	6.5%	450,510	6.5%
min DP \geq 3	Spy 94a	628,697	9.1%	628,007	9.1%
min DP \geq 3	Spy 94a (snpAD)	656,745	9.5%	654,847	9.5%
min DP \geq 3	Denisova 4	1,311,857	19.0%	1,311,543	19.0%
min DP \geq 3	Denisova 4 (snpAD)	1,370,096	19.8%	1,369,413	19.8%
min DP \geq 3	Denisova 8	3,902,126	56.4%	3,901,009	56.4%
min DP \geq 3	Denisova 8 (snpAD)	4,263,466	61.7%	4,261,828	61.7%
min DP \geq 3	Mezmaiskaya 2	6,141,494	88.8%	6,140,382	88.8%
min DP \geq 3	Mezmaiskaya 2 (snpAD)	6,458,882	93.4%	6,457,709	93.4%
min DP \geq 3 max DP \leq 98% quantile	Spy 94a (shotgun)	121,423	1.8%	121,391	1.8%
min DP \geq 3 max DP \leq 98% quantile	Mezmaiskaya 2 (shotgun)	425,654	6.2%	425,561	6.2%
min DP \geq 3 max DP \leq 98% quantile	Spy 94a	580,501	8.4%	579,836	8.4%

min DP \geq 3 max DP \leq 98% quantile	Spy 94a (snpAD)	605,324	8.8%	603,476	8.7%
min DP \geq 3 max DP \leq 98% quantile	Denisova 4	1,222,241	17.7%	1,221,944	17.7%
min DP \geq 3 max DP \leq 98% quantile	Denisova 4 (snpAD)	1,274,242	18.4%	1,273,579	18.4%
min DP \geq 3 max DP \leq 98% quantile	Denisova 8	3,809,899	55.1%	3,808,791	55.1%
min DP \geq 3 max DP \leq 98% quantile	Denisova 8 (snpAD)	4,167,667	60.3%	4,166,038	60.3%
min DP \geq 3 max DP \leq 98% quantile	Mezmaiskaya 2	6,038,340	87.4%	6,037,229	87.3%
min DP \geq 3 max DP \leq 98% quantile	Mezmaiskaya 2 (snpAD)	6,352,744	91.9%	6,351,572	91.9%

Table S8. Counts of sites for each archaic human Y chromosome in 6.9 Mb capture regions - raw counts, or counts of sites which passed the two filtering criteria used (first column). Multiple records for the same individual indicate different versions of the data (shotgun sequences as opposed to capture) or different ways of calling genotypes (consensus genotype calling or genotype calling using snpAD). Reported are numbers for all sites and for sites excluding C-T and G-A polymorphisms. The proportions are calculated relative to the total number of available sites (6,912,728).

filter	name	all SNPs		excluding C-to-T/G-to-A	
		count	proportion	count	proportion
unfiltered	El Sidrón 1253 (560 kb)	530,172	95.3%	530,046	95.3%
min DP \geq 3	El Sidrón 1253 (560 kb)	477,397	85.8%	477,293	85.8%
min DP \geq 3 max DP \leq 98% quantile	El Sidrón 1253 (560 kb)	467,563	84.1%	467,461	84.0%

Table S9. Counts of sites for the 560 kb capture regions - raw counts, or counts of sites which passed the two filtering criteria used (first column). Reported are numbers for all sites and for sites excluding C-to-T and G-to-A polymorphisms. The proportions are calculated relative to the total number of available sites (556,259).

individual	haplogroup
S_Finnish-2	I1a1b5
S_Sardinian-1	J1b2b
S_BedouinB-1	J1b2b
S_Punjabi-1	J2a1
S_French-1	J2a1b1
S_Turkish-1	R1b1a2a1a2c1g
S_Karitiana-1	Q1a2a1a1
S_Saami-2	N1c1a1a2a
S_Burmese-1	O3a1c2
S_Thai-1	O3a2c
S_Dai-2	O2a1
S_Han-2	O1a1
S_Papuan-2	S
S_Dinka-1	E2a
S_Gambian-1	E1b1a1a1f
S_Ju_hoan_North-1	B2b1b
S_Mandenka-1	E1b1a1a1g1
S_Mbuti-1	E1b1a1a1g1
S_Yoruba-2	E1b1a1a1f1b1
A00	A00
A00-1	A00
A00-2	A00

Table S10. Haplogroups of present-day human Y chromosomes used in our reference panel. Haplogroup names were taken from the SGDP annotation table (18). Haplogroups of the African panel are highlighted in gray. The A00 individual represents a merge of lower coverage sequences of two individuals, here named A00-1 and A00-2 (Table S4).

African	a	b	c	d	e	f	total	mutation rate (bp ⁻¹ × year ⁻¹)	TMRCA _{AFR} (years ago)
A00	1060.2	1.2	8.3	214.5	14.2	1103.1	6,064,477	7.34e-10	249211.63
A00-2	1054.5	1.2	8.2	214.7	14.6	1092.7	6,045,632	7.35e-10	247074.54
A00-1	1004.2	1.2	8.1	207.2	14.2	1054.1	5,855,414	7.32e-10	246756.41
S_Ju_hoan_North_1	261.5	2.7	8.5	216.3	14.5	439.2	6,100,839	7.35e-10	98443.62
S_Mbuti_1	116.5	1.2	8.9	217.5	14.5	344.7	6,123,504	7.37e-10	76782.21
S_Dinka_1	110.8	2.1	8.7	217.4	14.5	321.2	6,090,467	7.40e-10	71573.49
S_Mandenka_1	120.9	1.2	8.8	218.3	14.5	321.0	6,123,094	7.39e-10	71226.46
S_Gambian_1	118.8	1.2	8.5	218.1	14.5	320.2	6,096,582	7.42e-10	71127.40
S_Yoruba_2	118.9	1.2	8.7	217.6	13.5	320.9	6,121,558	7.41e-10	71119.46

Table S11. Branch counts and estimates of mutation rate and TMRCA between an African lineage and a panel of 13 non-African Y chromosomes. All quantities represent averages across all non-African Y chromosomes (Table S10). Counts in columns *a* to *f* represent counts of site patterns as shown in Figs. S11 and S12. “Total” represents the number of sites out of the total 6.9 Mb of target sequence available for the analysis. The last two columns represent the inferred mutation rate based on the *Ust’-Ishim* branch-shortening and the average TMRCA between a given African and a panel of non-Africans calculated from the length of the *a + d* branch as shown in Fig. S11. Bootstrap 95% confidence intervals for the A00-based mutation rate and split time in the first row (which we use to estimate archaic-modern human TMRCA) were estimated as [6.265082e-10, 8.464594e-10] mutations per bp per year and [213,342.15, 292,897.37] years ago, respectively.

name	a	b	c	d	e	f	total	TMRCAs [years ago]	TMRCAs (lower CI)	TMRCAs (upper CI)
Denisova 4	142.2	2.0	2.2	77.8	125.8	165.9	1,084,363.6	708,133.1	549,584.2	932,413.7
Denisova 8	583.5	8.8	8.5	318.0	410.4	583.9	3,372,262.6	706,874.9	607,066.1	834,698.5
Mezmaiskaya 2	301.8	13.8	16.4	625.1	609.1	651.9	5,349,303.5	369,637.7	326,438.0	420,072.7
Spy 94a	17.4	4.5	0.2	42.0	49.5	87.6	510,735.5	353,265.5	286,873.1	450,423.5
El Sidrón 1253	27.9	0.0	2.5	78.8	38.2	62.3	414,420.4	339,207.2	274,572.8	408,045.7

Table S12. Observed branch counts and estimates of TMRCAs between archaic and modern human Y chromosomes. All quantities represent averages across a panel of 13 non-African Y chromosomes (Table S10) and are based on A00-based estimates of mutation rate and $TMRCAs_{AFR}$. Counts in columns *a* to *f* represent counts of site patterns as shown in Figs. S14 and S15. “Total” represents the number of sites out of the total 6.9 Mb of target sequence available for the analysis.

REF count	ALT count	proportion REF	TMRC A	TMRC A (lower CI)	TMRC A (upper CI)	minimum coverage	calculated on
71	469	0.1315	92250.32	83814.88	102157.6	1	all sites
51	314	0.1397	94962.33	83315.75	107309.6	2	all sites
20	141	0.1242	89862.87	74453.89	108808.4	3	all sites
13	66	0.1646	103130.37	77184.06	129130.5	4	all sites
7	37	0.1591	101332.33	67250.74	141779.7	5	all sites
43	266	0.1392	94775.67	82515.76	108105.1	1	no C-T/G-A
32	173	0.1561	100347.68	83703.34	117665.3	2	no C-T/G-A
13	74	0.1494	98152.86	75463.15	125756.6	3	no C-T/G-A
7	33	0.1750	106565.56	69559.13	149511.3	4	no C-T/G-A
5	18	0.2174	120510.02	67229.10	180578.4	5	no C-T/G-A

Table S13. Point estimates and 95% C.I. for the TMRC A of *Mezmaiskaya 2* and *Spy 94a* Y chromosomes as shown in Fig. S26.

Table S14. DNA library, sequencing information and capture statistics. Table can be found online as an Excel file.

References and Notes

1. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Y. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspina, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušić, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the Neandertal genome. *Science* 328, 710–722 (2010). [doi:10.1126/science.1188021](https://doi.org/10.1126/science.1188021) [Medline](#)
2. M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338, 222–226 (2012). [doi:10.1126/science.1224344](https://doi.org/10.1126/science.1224344) [Medline](#)
3. S. Lippold, H. Xu, A. Ko, M. Li, G. Renaud, A. Butthof, R. Schröder, M. Stoneking, Human paternal and maternal demographic histories: Insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* 5, 13 (2014). [doi:10.1186/2041-2223-5-13](https://doi.org/10.1186/2041-2223-5-13) [Medline](#)
4. M. Karmin, L. Saag, M. Vicente, M. A. Wilson Sayres, M. Järve, U. G. Talas, S. Rootsi, A.-M. Ilumäe, R. Mägi, M. Mitt, L. Pagani, T. Puurand, Z. Faltyskova, F. Clemente, A. Cardona, E. Metspalu, H. Sahakyan, B. Yunusbayev, G. Hudjashov, M. DeGiorgio, E.-L. Loogväli, C. Eichstaedt, M. Eelmets, G. Chaubey, K. Tambets, S. Litvinov, M. Mormina, Y. Xue, Q. Ayub, G. Zoraqi, T. S. Korneliussen, F. Akhatova, J. Lachance, S. Tishkoff, K. Momynaliev, F.-X. Ricaut, P. Kusuma, H. Razafindrazaka, D. Pierron, M. P. Cox, G. N. N. Sultana, R. Willerslev, C. Muller, M. Westaway, D. Lambert, V. Skaro, L. Kovačević, S. Turdikulova, D. Dalimova, R. Khusainova, N. Trofimova, V. Akhmetova, I. Khidiyatova, D. V. Lichman, J. Isakova, E. Pocheshkhova, Z. Sabitov, N. A. Barashkov, P. Nymadawa, E. Mihailov, J. W. T. Seng, I. Evseeva, A. B. Migliano, S. Abdullah, G. Andriadze, D. Primorac, L. Atramentova, O. Utevska, L. Yepiskoposyan, D. Marjanović, A. Kushniarevich, D. M. Behar, C. Gilissen, L. Vissers, J. A. Veltman, E. Balanovska, M. Derenko, B. Malyarchuk, A. Metspalu, S. Fedorova, A. Eriksson, A. Manica, F. L. Mendez, T. M. Karafet, K. R. Veeramah, N. Bradman, M. F. Hammer, L. P. Osipova, O. Balanovsky, E. K. Khusnutdinova, K. Johnsen, M. Remm, M. G. Thomas, C. Tyler-Smith, P. A. Underhill, E. Willerslev, R. Nielsen, M. Metspalu, R. Villems, T. Kivisild, A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* 25, 459–466 (2015). [doi:10.1101/gr.186684.114](https://doi.org/10.1101/gr.186684.114) [Medline](#)
5. I. Olalde, S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, M. Silva, K. Dulias, C. J. Edwards, F. Gandini, M. Pala, P. Soares, M. Ferrando-Bernal, N. Adamski, N. Broomandkoshbacht, O. Cheronet, B. J. Culleton, D. Fernandes, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, Z. Zhang, J. M. Jiménez Arenas, I. J. Toro

- Moyano, D. C. Salazar-García, P. Castanyer, M. Santos, J. Tremoleda, M. Lozano, P. García Borja, J. Fernández-Eraso, J. A. Mujika-Alustiza, C. Barroso, F. J. Bermúdez, E. Viguera Mínguez, J. Burch, N. Coromina, D. Vivó, A. Cebrià, J. M. Fullola, O. García-Puchol, J. I. Morales, F. X. Oms, T. Majó, J. M. Vergès, A. Díaz-Carvajal, I. Ollich-Castanyer, F. J. López-Cachero, A. M. Silva, C. Alonso-Fernández, G. Delibes de Castro, J. Jiménez Echevarría, A. Moreno-Márquez, G. Pascual Berlanga, P. Ramos-García, J. Ramos-Muñoz, E. Vijande Vila, G. Aguilera Arzo, Á. Esparza Arroyo, K. T. Lillios, J. Mack, J. Velasco-Vázquez, A. Waterman, L. Benítez de Lugo Enrich, M. Benito Sánchez, B. Agustí, F. Codina, G. de Prado, A. Estalrich, Á. Fernández Flores, C. Finlayson, G. Finlayson, S. Finlayson, F. Giles-Guzmán, A. Rosas, V. Barciela González, G. García Atiénzar, M. S. Hernández Pérez, A. Llanos, Y. Carrión Marco, I. Collado Beneyto, D. López-Serrano, M. Sanz Tormo, A. C. Valera, C. Blasco, C. Liesau, P. Ríos, J. Daura, M. J. de Pedro Michó, A. A. Diez-Castillo, R. Flores Fernández, J. Francés Farré, R. Garrido-Pena, V. S. Gonçalves, E. Guerra-Doce, A. M. Herrero-Corral, J. Juan-Cabanilles, D. López-Reyes, S. B. McClure, M. Merino Pérez, A. Oliver Foix, M. Sanz Borràs, A. C. Sousa, J. M. Vidal Encinas, D. J. Kennett, M. B. Richards, K. Werner Alt, W. Haak, R. Pinhasi, C. Lalueza-Fox, D. Reich, The genomic history of the Iberian Peninsula over the past 8000 years. *Science* 363, 1230–1234 (2019).
[doi:10.1126/science.aav4040](https://doi.org/10.1126/science.aav4040) [Medline](#)
6. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49 (2014).
[doi:10.1038/nature12886](https://doi.org/10.1038/nature12886) [Medline](#)
7. C. Posth, C. Wißing, K. Kitagawa, L. Pagani, L. van Holstein, F. Racimo, K. Wehrberger, N. J. Conard, C. J. Kind, H. Bocherens, J. Krause, Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat. Commun.* 8, 16046 (2017). [doi:10.1038/ncomms16046](https://doi.org/10.1038/ncomms16046) [Medline](#)
8. M. Meyer, Q. Fu, A. Aximu-Petri, I. Glocke, B. Nickel, J.-L. Arsuaga, I. Martínez, A. Gracia, J. M. B. de Castro, E. Carbonell, S. Pääbo, A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 505, 403–406 (2014).
[doi:10.1038/nature12788](https://doi.org/10.1038/nature12788) [Medline](#)
9. M. Meyer, J.-L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, B. Nickel, I. Martínez, A. Gracia, J. M. Bermúdez de Castro, E. Carbonell, B. Viola, J. Kelso, K. Prüfer, S. Pääbo, Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* 531, 504–507 (2016). [doi:10.1038/nature17405](https://doi.org/10.1038/nature17405) [Medline](#)
10. F. L. Mendez, G. D. Poznik, S. Castellano, C. D. Bustamante, The divergence of Neandertal and modern human Y chromosomes. *Am. J. Hum. Genet.* 98, 728–734 (2016).
[doi:10.1016/j.ajhg.2016.02.023](https://doi.org/10.1016/j.ajhg.2016.02.023) [Medline](#)

11. S. Sawyer, G. Renaud, B. Viola, J.-J. Hublin, M.-T. Gansauge, M. V. Shunkov, A. P. Derevianko, K. Prüfer, J. Kelso, S. Pääbo, Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15696–15700 (2015). [doi:10.1073/pnas.1519905112](https://doi.org/10.1073/pnas.1519905112) [Medline](#)
12. K. Douka, V. Slon, Z. Jacobs, C. B. Ramsey, M. V. Shunkov, A. P. Derevianko, F. Mafessoni, M. B. Kozlikin, B. Li, R. Grün, D. Comeskey, T. Devièse, S. Brown, B. Viola, L. Kinsley, M. Buckley, M. Meyer, R. G. Roberts, S. Pääbo, J. Kelso, T. Higham, Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature* 565, 640–644 (2019). [doi:10.1038/s41586-018-0870-z](https://doi.org/10.1038/s41586-018-0870-z) [Medline](#)
13. M. Hajdinjak, Q. Fu, A. Hübner, M. Petr, F. Mafessoni, S. Grote, P. Skoglund, V. Narasimham, H. Rougier, I. Crevecoeur, P. Semal, M. Soressi, S. Talamo, J.-J. Hublin, I. Gušić, Ž. Kučan, P. Rudan, L. V. Golovanova, V. B. Doronichev, C. Posth, J. Krause, P. Korlević, S. Nagel, B. Nickel, M. Slatkin, N. Patterson, D. Reich, K. Prüfer, M. Meyer, S. Pääbo, J. Kelso, Reconstructing the genetic history of late Neanderthals. *Nature* 555, 652–656 (2018). [doi:10.1038/nature26151](https://doi.org/10.1038/nature26151) [Medline](#)
14. See supplementary materials.
15. R. E. Wood, T. F. G. Higham, T. De Torres, N. Tisnérat-Laborde, H. Valladas, J. E. Ortiz, C. Lalueza-Fox, S. Sánchez-Moral, J. C. Cañaveras, A. Rosas, D. Santamaría, M. De La Rasilla, A new date for the Neanderthals from El Sidrón Cave (Asturias, northern Spain). *Archaeometry* 55, 148–158 (2013). [doi:10.1111/j.1475-4754.2012.00671.x](https://doi.org/10.1111/j.1475-4754.2012.00671.x)
16. S. Castellano, G. Parra, F. A. Sánchez-Quinto, F. Racimo, M. Kuhlwilm, M. Kircher, S. Sawyer, Q. Fu, A. Heinze, B. Nickel, J. Dabney, M. Siebauer, L. White, H. A. Burbano, G. Renaud, U. Stenzel, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Rudan, D. Brajković, Ž. Kucan, I. Gušić, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Meyer, J. Kelso, A. M. Andrés, S. Pääbo, Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6666–6671 (2014). [doi:10.1073/pnas.1405138111](https://doi.org/10.1073/pnas.1405138111) [Medline](#)
17. Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-García, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B. Viola, S. Pääbo, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514, 445–449 (2014). [doi:10.1038/nature13810](https://doi.org/10.1038/nature13810) [Medline](#)
18. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T.

- Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206 (2016). [doi:10.1038/nature18964](https://doi.org/10.1038/nature18964) [Medline](#)
19. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358, 655–658 (2017). [doi:10.1126/science.aao1887](https://doi.org/10.1126/science.aao1887) [Medline](#)
20. M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, H. A. Burbano, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušić, T. Marques-Bonet, A. M. Andrés, B. Viola, S. Pääbo, M. Meyer, A. Siepel, S. Castellano, Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* 530, 429–433 (2016). [doi:10.1038/nature16544](https://doi.org/10.1038/nature16544) [Medline](#)
21. M. J. Hubisz, A. L. Williams, A. Siepel, Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLOS Genet.* 16, e1008895 (2020). [doi:10.1371/journal.pgen.1008895](https://doi.org/10.1371/journal.pgen.1008895) [Medline](#)
22. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (The Blackburn Press, 1970).
23. K. Harris, R. Nielsen, The genetic cost of Neanderthal introgression. *Genetics* 203, 881–891 (2016). [doi:10.1534/genetics.116.186890](https://doi.org/10.1534/genetics.116.186890) [Medline](#)
24. I. Juric, S. Aeschbacher, G. Coop, The strength of selection against neanderthal introgression. *PLOS Genet.* 12, e1006340 (2016). [doi:10.1371/journal.pgen.1006340](https://doi.org/10.1371/journal.pgen.1006340) [Medline](#)
25. B. C. Haller, P. W. Messer, SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.* 36, 632–637 (2019). [doi:10.1093/molbev/msy228](https://doi.org/10.1093/molbev/msy228) [Medline](#)
26. S. Colaco, D. Modi, Genetics of the human Y chromosome and its association with male infertility. *Reprod. Biol. Endocrinol.* 16, 14 (2018). [doi:10.1186/s12958-018-0330-5](https://doi.org/10.1186/s12958-018-0330-5) [Medline](#)
27. J. W. Ballard, M. C. Whitlock, The incomplete natural history of mitochondria. *Mol. Ecol.* 13, 729–744 (2004). [doi:10.1046/j.1365-294X.2003.02063.x](https://doi.org/10.1046/j.1365-294X.2003.02063.x) [Medline](#)
28. T. Bonnet, R. Leblois, F. Rousset, P.-A. Crochet, A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution* 71, 2140–2158 (2017). [doi:10.1111/evo.13296](https://doi.org/10.1111/evo.13296) [Medline](#)
29. F. A. Seixas, P. Boursot, J. Melo-Ferreira, The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biol.* 19, 91 (2018). [doi:10.1186/s13059-018-1471-8](https://doi.org/10.1186/s13059-018-1471-8) [Medline](#)

30. L. Skov, M. H. Schierup; Danish Pan Genome Consortium, Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLOS Genet.* 13, e1006834 (2017). [doi:10.1371/journal.pgen.1006834](https://doi.org/10.1371/journal.pgen.1006834) [Medline](#)
31. M. Petr, “The evolutionary history of Neandertal and Denisovan Y chromosomes” - code and Jupyter notebooks. Zenodo (2020); [doi:10.5281/zenodo.3941654](https://doi.org/10.5281/zenodo.3941654).
32. M. Petr, “The evolutionary history of Neandertal and Denisovan Y chromosomes” - Y chromosome capture designs. Zenodo (2020); [doi: 10.5281/zenodo.3940568](https://doi.org/10.5281/zenodo.3940568).
33. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580 (1999). [doi:10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573) [Medline](#)
34. Q. Fu, M. Meyer, X. Gao, U. Stenzel, H. A. Burbano, J. Kelso, S. Pääbo, DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2223–2227 (2013). [doi:10.1073/pnas.1221359110](https://doi.org/10.1073/pnas.1221359110) [Medline](#)
35. D. I. Cruz-Dávalos, M. A. Nieves-Colón, A. Sockell, G. D. Poznik, H. Schroeder, A. C. Stone, C. D. Bustamante, A.-S. Malaspinas, M. C. Ávila-Arcos, In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics* 19, 608 (2018). [doi:10.1186/s12864-018-4945-x](https://doi.org/10.1186/s12864-018-4945-x) [Medline](#)
36. J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, M. Meyer, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763 (2013). [doi:10.1073/pnas.1314445110](https://doi.org/10.1073/pnas.1314445110) [Medline](#)
37. P. Korlević, T. Gerber, M.-T. Gansauge, M. Hajdinjak, S. Nagel, A. Aximu-Petri, M. Meyer, Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* 59, 87–93 (2015). [doi:10.2144/000114320](https://doi.org/10.2144/000114320) [Medline](#)
38. I. Glocke, M. Meyer, Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* 27, 1230–1237 (2017). [doi:10.1101/gr.219675.116](https://doi.org/10.1101/gr.219675.116) [Medline](#)
39. N. Rohland, I. Glocke, A. Aximu-Petri, M. Meyer, Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat. Protoc.* 13, 2447–2461 (2018). [doi:10.1038/s41596-018-0050-5](https://doi.org/10.1038/s41596-018-0050-5) [Medline](#)
40. M. T. Gansauge, M. Meyer, Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* 8, 737–748 (2013). [doi:10.1038/nprot.2013.038](https://doi.org/10.1038/nprot.2013.038) [Medline](#)
41. M.-T. Gansauge, T. Gerber, I. Glocke, P. Korlević, L. Lippik, S. Nagel, L. M. Riehl, A. Schmidt, M. Meyer, Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45, e79 (2017). [doi:10.1093/nar/gkx033](https://doi.org/10.1093/nar/gkx033) [Medline](#)
42. V. Slon, B. Viola, G. Renaud, M.-T. Gansauge, S. Benazzi, S. Sawyer, J.-J. Hublin, M. V. Shunkov, A. P. Derevianko, J. Kelso, K. Prüfer, M. Meyer, S. Pääbo, A fourth Denisovan individual. *Sci. Adv.* 3, e1700186 (2017). [doi:10.1126/sciadv.1700186](https://doi.org/10.1126/sciadv.1700186) [Medline](#)

43. M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40, e3 (2012). [doi:10.1093/nar/gkr771](https://doi.org/10.1093/nar/gkr771) [Medline](#)
44. J. Dabney, M. Meyer, Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52, 87–94 (2012). [doi:10.2144/000113809](https://doi.org/10.2144/000113809) [Medline](#)
45. L. Bokelmann, M. Hajdinjak, S. Peyrégne, S. Brace, E. Essel, C. de Filippo, I. Glocke, S. Grote, F. Mafessoni, S. Nagel, J. Kelso, K. Prüfer, B. Vernot, I. Barnes, S. Pääbo, M. Meyer, C. Stringer, A genetic analysis of the Gibraltar Neanderthals. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15610–15615 (2019). [doi:10.1073/pnas.1903984116](https://doi.org/10.1073/pnas.1903984116) [Medline](#)
46. G. Renaud, U. Stenzel, J. Kelso, leeHom: Adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* 42, e141 (2014). [doi:10.1093/nar/gku699](https://doi.org/10.1093/nar/gku699) [Medline](#)
47. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009). [doi:10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) [Medline](#)
48. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009). [doi:10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) [Medline](#)
49. F. L. Mendez, T. Krahn, B. Schrack, A.-M. Krahn, K. R. Veeramah, A. E. Woerner, F. L. M. Fomine, N. Bradman, M. G. Thomas, T. M. Karafet, M. F. Hammer, An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am. J. Hum. Genet.* 92, 454–459 (2013). [doi:10.1016/j.ajhg.2013.02.002](https://doi.org/10.1016/j.ajhg.2013.02.002) [Medline](#)
50. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010). [doi:10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033) [Medline](#)
51. A. W. Briggs, U. Stenzel, P. L. F. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, S. Pääbo, Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14616–14621 (2007). [doi:10.1073/pnas.0704665104](https://doi.org/10.1073/pnas.0704665104) [Medline](#)
52. T. Günther, C. Nettelblad, The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genet.* 15, e1008302 (2019). [doi:10.1371/journal.pgen.1008302](https://doi.org/10.1371/journal.pgen.1008302) [Medline](#)
53. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993 (2011). [doi:10.1093/bioinformatics/btr509](https://doi.org/10.1093/bioinformatics/btr509) [Medline](#)
54. K. Prüfer, snpAD: An ancient DNA genotype caller. *Bioinformatics* 34, 4165–4171 (2018). [doi:10.1093/bioinformatics/bty507](https://doi.org/10.1093/bioinformatics/bty507) [Medline](#)
55. T. Kivisild, The study of human Y chromosome variation through ancient DNA. *Hum. Genet.* 136, 529–546 (2017). [doi:10.1007/s00439-017-1773-z](https://doi.org/10.1007/s00439-017-1773-z) [Medline](#)
56. K. P. Schliep, phangorn: Phylogenetic analysis in R. *Bioinformatics* 27, 592–593 (2011). [doi:10.1093/bioinformatics/btq706](https://doi.org/10.1093/bioinformatics/btq706) [Medline](#)

-
57. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528 (2019). [doi:10.1093/bioinformatics/bty633](https://doi.org/10.1093/bioinformatics/bty633) [Medline](#)
58. M. K. Kuhner, J. Felsenstein, A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468 (1994). [Medline](#)
59. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36 (2017). [doi:10.1111/2041-210X.12628](https://doi.org/10.1111/2041-210X.12628)
60. P. Hallast, C. Batini, D. Zadik, P. Maisano Delser, J. H. Wetton, E. Arroyo-Pardo, G. L. Cavalleri, P. de Knijff, G. Destro Bisol, B. M. Dupuy, H. A. Eriksen, L. B. Jorde, T. E. King, M. H. Larmuseau, A. López de Munain, A. M. López-Parra, A. Loutradis, J. Milasin, A. Novelletto, H. Pamjav, A. Sajantila, W. Schempp, M. Sears, A. Tolun, C. Tyler-Smith, A. Van Geystelen, S. Watkins, B. Winney, M. A. Jobling, The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol. Biol. Evol.* 32, 661–673 (2015). [doi:10.1093/molbev/msu327](https://doi.org/10.1093/molbev/msu327) [Medline](#)
61. C. Barbieri, A. Hübner, E. Macholdt, S. Ni, S. Lippold, R. Schröder, S. W. Mpoloka, J. Purps, L. Roewer, M. Stoneking, B. Pakendorf, Refining the Y chromosome phylogeny with southern African sequences. *Hum. Genet.* 135, 541–553 (2016). [doi:10.1007/s00439-016-1651-0](https://doi.org/10.1007/s00439-016-1651-0) [Medline](#)
62. P. Hallast, P. Maisano Delser, C. Batini, D. Zadik, M. Rocchi, W. Schempp, C. Tyler-Smith, M. A. Jobling, Great ape Y chromosome and mitochondrial DNA phylogenies reflect subspecies structure and patterns of mating and dispersal. *Genome Res.* 26, 427–439 (2016). [doi:10.1101/gr.198754.115](https://doi.org/10.1101/gr.198754.115) [Medline](#)
63. M. Petr, S. Pääbo, J. Kelso, B. Vernot, Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1639–1644 (2019). [doi:10.1073/pnas.1814338116](https://doi.org/10.1073/pnas.1814338116) [Medline](#)
64. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell* 180, 677–687.e16 (2020). [doi:10.1016/j.cell.2020.01.012](https://doi.org/10.1016/j.cell.2020.01.012) [Medline](#)
65. A. Eyre-Walker, M. Woolfit, T. Phelps, The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173, 891–900 (2006). [doi:10.1534/genetics.106.057570](https://doi.org/10.1534/genetics.106.057570) [Medline](#)
66. A. R. Boyko, S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark, C. D. Bustamante, Assessing the evolutionary impact of amino acid mutations in the human genome. *PLOS Genet.* 4, e1000083 (2008). [doi:10.1371/journal.pgen.1000083](https://doi.org/10.1371/journal.pgen.1000083) [Medline](#)
67. W. A. Brashear, T. Raudsepp, W. J. Murphy, Evolutionary conservation of Y Chromosome ampliconic gene families despite extensive structural variation. *Genome Res.* 28, 1841–1851 (2018). [doi:10.1101/gr.237586.118](https://doi.org/10.1101/gr.237586.118) [Medline](#)

Chapter 3

admixr—R package for reproducible analyses using
ADMIXTOOLS

Petr, M., Vernot, B., Kelso, J.

Bioinformatics, 35, 3194–3195 (2019)

Genetics and population analysis

admixr—R package for reproducible analyses using ADMIXTOOLS

Martin Petr *, Benjamin Vernot  and Janet Kelso *

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 4, 2018; editorial decision on January 5, 2019; accepted on January 17, 2019

Abstract

Summary: We present a new R package *admixr*, which provides a convenient interface for performing reproducible population genetic analyses (f_3 , D , f_4 , f_4 -ratio, *qpWave* and *qpAdm*), as implemented by command-line programs in the ADMIXTOOLS software suite. In a traditional ADMIXTOOLS workflow, the user must first generate a set of text configuration files tailored to each individual analysis, often using a combination of shell scripting and manual text editing. The non-tabular output files then need to be parsed to extract values of interest prior to further analyses. Our package simplifies this process by automating all low-level configuration and parsing steps, making analyses as simple as running a single R command. Furthermore, we provide a set of R functions for processing, filtering and manipulating datasets in the EIGENSTRAT format. By unifying all steps of the workflow under a single R framework, this package enables the automation of analytic pipelines, significantly improving the reproducibility of population genetic studies.

Availability and implementation: The source code of the R package is available under the MIT license. Installation instructions, reference manual and a tutorial can be found on the package website at <https://bioinf.eva.mpg.de/admixr>.

Contact: mp@bodkan.net or kelso@eva.mpg.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The growing number of ancient and modern genome sequences have transformed our understanding of the evolutionary history of humans and other species. Several statistical methods have been developed to make inferences about past population movements and admixtures from genomic data. Chief among these has been a series of population genetic methods (D , f_3 , f_4 , f_4 -ratio, *qpWave* and *qpAdm*) for estimating the amounts of genetic drift shared between populations, testing admixture hypotheses and estimating admixture proportions, implemented as command-line utilities in the ADMIXTOOLS software suite (Patterson *et al.*, 2012). Although ADMIXTOOLS has been used in many recent studies of human ancient DNA (Fu *et al.*, 2016; Haak *et al.*, 2015; Hajdinjak *et al.*, 2018; Lazaridis *et al.*, 2016), the tools in this package are rather cumbersome to use. First, each individual analysis or hypothesis test relies on a set of configuration files, which have to be generated

using a combination of shell scripting and manual editing. Second, after running an ADMIXTOOLS command on the command-line, the user needs to extract relevant values from a non-tabular text file before they can be imported into software such as R for further analysis and plotting. This workflow is slow and potentially error-prone, especially if the user wishes to quickly iterate through different hypotheses involving many different populations or samples. Most importantly, however, it makes it challenging to conduct fully reproducible research. To overcome these challenges, we present a new R package for population admixture analyses which utilizes the ADMIXTOOLS software suite for the underlying calculations, but that provides a unified and convenient R interface. The package completely automates the generation, processing and parsing of all intermediate files, hiding all low-level details from the user, and allowing them to focus on the analysis itself. Importantly, unifying the entire analytic workflow in a single environment makes it

possible to implement and share fully automated, reproducible analytic pipelines.

2 Implementation

The *admixr* package is implemented using the R programming language. It consists of several wrapper functions (calling ADMIXTOOLS commands internally from R), and a set of complementary functions for filtering and processing datasets in the EIGENSTRAT file format required by ADMIXTOOLS (Patterson *et al.*, 2012).

An EIGENSTRAT dataset is represented by an S3 object of the class EIGENSTRAT, which is created using the `eigenstrat()` constructor function, and encapsulates the paths to a trio of 'ind', 'snp' and 'geno' files:

```
> snps <- eigenstrat("~/path/to/eigenstrat/data")
> snps
EIGENSTRAT object
=====
components:
  ind file: ~/path/to/eigenstrat/data.ind
  snp file: ~/path/to/eigenstrat/data.snp
  geno file: ~/path/to/eigenstrat/data.geno
```

All other functions in the package accept this object as their first argument, and perform either a requested calculation on it (returning an R data frame for further analysis), or return a new, modified EIGENSTRAT S3 object (in case of filtering and processing functions) which can be used in additional downstream steps or calculations.

The core functionality of the package consists of the following set of R functions: `f3()`, `d()`, `f4()`, `f4ratio()`, `qpWave()` and `qpAdm()`, each implemented as a wrapper around one of the command-line programs distributed as part of the ADMIXTOOLS package.

3 Example usage

Performing even the most trivial analysis using ADMIXTOOLS presents a significant amount of overhead for the user. For example, to estimate the proportion of Neandertal ancestry in a set of individuals, X, the user would typically calculate an f_4 -ratio statistic such as:

$$\frac{f_4(\text{Altai, Chimp; X, Mbuti})}{f_4(\text{Altai, Chimp; Vindija, Mbuti})} \quad (1)$$

The user first needs to create a file with a list of samples in each position of both f_4 statistics, a parameter file specifying the paths to a trio of EIGENSTRAT component files, then manually run the `qpF4ratio` command-line program, and then capture and parse its output to obtain relevant values (see [Supplementary Information](#) for a complete example workflow using a traditional ADMIXTOOLS approach). Note that changing the analysis setup [such as including a different set of populations in [Equation \(1\)](#)], performing the analysis on a subset of the genome, or modifying the analysis in another way, requires changes to be made to its configuration files. This presents a significant overhead for the user, especially when iterating through a complex set of population genetic hypotheses.

In contrast, using the *admixr* package, the same analysis can be performed with just the following snippet of R code:

```
result <- f4ratio(
  X = c("French", "Han", "Papuan"),
  A = "Altai", B = "Vindija", C = "Mbuti", O =
    "Chimp",
  data = eigenstrat("<path to EIGENSTRAT data>")
)
```

Internally, the `f4ratio()` function performs all configuration and parsing work, and returns an R data frame which can be immediately used for further statistical analysis and plotting:

```
> result
  A      B      X      C      O  alpha  stderr  Zscore
Altai Vindija French Mbuti Chimp 0.019696 0.003114 6.324
Altai Vindija   Han Mbuti Chimp 0.024379 0.003364 7.248
Altai Vindija Papuan Mbuti Chimp 0.032167 0.003499 9.193
```

All other *admixr* wrapper functions have a similar interface and are described in the tutorial vignette on the package website in more detail.

4 Additional functionality

The fact that ADMIXTOOLS requires the data to be in EIGENSTRAT format presents additional challenges for quality control, processing and filtering, as this format is not supported by standard bioinformatics tools. Our R package therefore provides additional functionality to simplify the processing and filtering of EIGENSTRAT genotype data. This includes:

- Reading and writing of ind, snp and geno file components.
- Filtering of SNPs based on regions specified in a BED file.
- Restricting analyses to sites carrying transversion SNPs.
- Renaming samples or grouping them into larger population groups.
- Merging of EIGENSTRAT datasets.
- Counting the number of sites present or missing in each sample.

Acknowledgements

We would like to thank Benjamin Peter and Hassan Shafiey for helpful comments, suggestions and testing.

Funding

This work was supported by funding from the Max Planck Society.

Conflict of Interest: none declared.

References

- Fu, Q. *et al.* (2016) The genetic history of Ice Age Europe. *Nature*, **534**, 200–205.
- Haak, W. *et al.* (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, **522**, 207–211.
- Hajdinjak, M. *et al.* (2018) Reconstructing the genetic history of late Neanderthals. *Nature*, **555**, 652–656.
- Lazaridis, I. *et al.* (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature*, **536**, 419–424.
- Patterson, N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Supplementary Information

All files used in this document and its source in R Markdown can be found at:
<https://github.com/bodkan/admixr-SI>.

A complete description of the functionality of the *admixr* package can be found in the tutorial vignette at: <https://bodkan.net/admixr/articles/tutorial.html>. In this supplement we provide an example analysis to demonstrate more clearly the distinctions between the traditional ADMIXTOOLS workflow and the *admixr* approach.

Example analysis: Estimating the proportions of Neandertal ancestry

Our goal is to estimate the proportions of Neandertal ancestry in a set of present-day European individuals (French, Sardinian and Czech) using the following f_4 -ratio statistic:

$$\alpha = \frac{f_4(\text{Altai Neandertal, Chimpanzee; European X, African})}{f_4(\text{Altai Neandertal, Chimpanzee; Vindija Neandertal, African})}$$

a) A traditional ADMIXTOOLS workflow

To perform this task using a traditional ADMIXTOOLS workflow, we have to do the following:

1. Create a pop file (`f4ratio.pop`).

This file specifies the order of populations in the f_4 statistic calculation above, and will contain the following (one f_4 -ratio setup per row, as required by ADMIXTOOLS):

```
Altai Chimp : French Yoruba :: Altai Chimp : Vindija Yoruba
Altai Chimp : Sardinian Yoruba :: Altai Chimp : Vindija Yoruba
Altai Chimp : Czech Yoruba :: Altai Chimp : Vindija Yoruba
```

2. Create a par file (`f4ratio_all.par`).

This is the main configuration file and contains the paths to a trio of EIGENSTRAT components and a path to the pop file created in the previous step.

```
genotypename: ./data/snps.geno
snpname: ./data/snps.snp
indivname: ./data/snps.ind
popfilename: ./f4ratio.pop
```

3. Run the `qpF4ratio` command-line program and capture its output:

```
qpF4ratio -p f4ratio_all.par > f4ratio_all.log
```

4. Inspect the output file (f4ratio_all.log) and extract the relevant values.

To extract the values of interest (on lines beginning with “result:”), we could either copy-paste them manually into Excel, or extract them using a combination of `grep/awk/sed` and save them into a text file for further analysis (we will omit these steps here).

```
qpF4ratio: parameter file: f4ratio_all.par
### THE INPUT PARAMETERS
##PARAMETER NAME: VALUE
genotypename: ./data/snps.geno
snpname: ./data/snps.snp
indivname: ./data/snps.ind
popfilename: ./f4ratio.pop
## qpF4ratio version: 310
nplist: 3
  0           Altai    1
  1           Chimp    1
  2           French   1
  3           Yoruba   1
  4           Vindija  1
  5           Sardinian 1
  6           Czech    1
jackknife block size: 0.050
snps: 2055980 indivs: 7
number of blocks for block jackknife: 557

alpha      std. err  Z (null=0)
result:    Altai      Chimp      French     Yoruba :    Altai      Chimp
Vindija    Yoruba      0.021646   0.003738   5.791 9d    Altai      Chimp
result:    Altai      Chimp      Sardinian  Yoruba :    Altai      Chimp
Vindija    Yoruba      0.025142   0.003594   6.996 9d    Altai      Chimp
result:    Altai      Chimp      Czech      Yoruba :    Altai      Chimp
Vindija    Yoruba      0.026637   0.003605   7.389 9d
## end of run
```

b) A new workflow using the *admixr* package

Running the following R code will perform all the low-level configuration and parsing work (steps 1-4 above) automatically, returning a simple R data frame:

```
> library(admixr)
>
> result <- f4ratio(
>   X = c("French", "Sardinian", "Czech"),
>   A = "Altai", B = "Vindija", C = "Yoruba", O = "Chimp",
>   data = eigenstrat("data/snps")
> )
```

```
> result
# A tibble: 3 x 8
  A      B      X      C      O      alpha stderr Zscore
  <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>
1 Altai Vindija French Yoruba Chimp 0.0216 0.00374 5.79
2 Altai Vindija Sardinian Yoruba Chimp 0.0251 0.00359 7.00
3 Altai Vindija Czech Yoruba Chimp 0.0266 0.00360 7.39
```

Estimate Neandertal ancestry proportions on a subset of the data.

Let's investigate a variation of the analysis above, this time estimating Neandertal ancestry proportions in a subset of the genome (in a set of regions specified in a BED file `regions.bed`).

a) A traditional ADMIXTOOLS workflow

All ADMIXTOOLS commands accept a parameter `badsnpname`, which can be specified in a `par` file and instructs ADMIXTOOLS to ignore SNPs at specified coordinates.

In a traditional ADMIXTOOLS workflow, we would have to create this file using some form of shell scripting. For example, we could do the following:

1. Convert the EIGENSTRAT `snp` file into a BED format:

```
awk -v OFS='\t' '{print $2, $4 - 1, $4, $0}' data/snps.snp > snps.bed
```

2. Intersect the newly generated BED file with coordinates of regions in `regions.bed`. We want to keep only those sites that fall outside of those regions and saving them in a `snp` format again.

```
bedtools intersect -v -a snps.bed -b regions.bed \
| cut -f 4- \
> excluded_sites.snp
```

3. Create a new `par` file (`f4ratio_subset.par`), with a new parameter `badsnpname`:

```
genotypename: ./data/snps.geno
snpname: ./data/snps.snp
indivname: ./data/snps.ind
popfilename: ./f4ratio.pop
badsnpname: ./excluded_sites.snp
```

4. Run `qpF4ratio` with the new `par` file and capture its output:

```
qpF4ratio -p f4ratio_subset.par > f4ratio_subset.log
```

5. Extract the relevant values from the output file (f4ratio_subset.log):

```
qpF4ratio: parameter file: f4ratio_subset.par
### THE INPUT PARAMETERS
##PARAMETER NAME: VALUE
genotypename: ./data/snps.geno
snpname: ./data/snps.snp
indivname: ./data/snps.ind
popfilename: ./f4ratio.pop
badsnpname: ./excluded_sites.snp
## qpF4ratio version: 310
nplist: 3
  0           Altai    1
  1           Chimp    1
  2           French   1
  3           Yoruba   1
  4           Vindija  1
  5           Sardinian 1
  6           Czech    1
jackknife block size: 0.050
snps: 1716304 indivs: 7
number of blocks for block jackknife: 556

alpha      std. err  Z (null=0)
result:    Altai     Chimp     French    Yoruba   :    Altai     Chimp
Vindija    Yoruba    0.023227  0.003882  5.984 9d
result:    Altai     Chimp    Sardinian  Yoruba   :    Altai     Chimp
Vindija    Yoruba    0.027174  0.003799  7.154 9d
result:    Altai     Chimp     Czech     Yoruba   :    Altai     Chimp
Vindija    Yoruba    0.027380  0.003821  7.165 9d
## end of run
```

b) A new workflow using the *admixr* package

Using the *admixr* package, we can use the function `filter_bed()` to do all filtering automatically:

```
> library(admixr)
>
> snps <- eigenstrat("data/snps")
> subset <- filter_bed(snps, "regions.bed") # create a new EIGENSTRAT object
>
> result <- f4ratio(
>   X = c("French", "Sardinian", "Czech"),
>   A = "Altai", B = "Vindija", C = "Yoruba", O = "Chimp",
>   data = subset
> )
```

```

> result

# A tibble: 3 x 8
  A      B      X      C      O      alpha  stderr  Zscore
  <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>
1 Altai Vindija French  Yoruba Chimp 0.0232 0.00388 5.98
2 Altai Vindija Sardinian Yoruba Chimp 0.0272 0.00380 7.15
3 Altai Vindija Czech    Yoruba Chimp 0.0274 0.00382 7.16

```

Furthermore, because all *admixr* functions are inspired by the “tidyverse philosophy”, we can utilize the `%>%` pipe operator from the *magrittr* package and run the whole analysis in one step:

```

> library(admixr)
> library(magrittr)
>
> result <-
>   eigenstrat("data/snps") %>%
>   filter_bed("regions.bed") %>%
>   f4ratio(
>     X = c("French", "Sardinian", "Czech"),
>     A = "Altai", B = "Vindija", C = "Yoruba", O = "Chimp"
>   )
>
> result

# A tibble: 3 x 8
  A      B      X      C      O      alpha  stderr  Zscore
  <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>
1 Altai Vindija French  Yoruba Chimp 0.0232 0.00388 5.98
2 Altai Vindija Sardinian Yoruba Chimp 0.0272 0.00380 7.15
3 Altai Vindija Czech    Yoruba Chimp 0.0274 0.00382 7.16

```

A more complex example

The “pipe-friendly” nature of *admixr* makes it possible to compose even more complex pipelines, such as the following:

```

> library(admixr)
> library(magrittr)
>
> result <-
>   eigenstrat("data/snps") %>%
>   relabel(Europeans = c("French", "Sardinian", "Czech")) %>%
>   filter_bed("regions.bed") %>%
>   transversions_only() %>%
>   f4ratio(
>     X = "Europeans",
>     A = "Altai", B = "Vindija", C = "Yoruba", O = "Chimp"
>   )

```



```
> result
```

```
# A tibble: 1 x 8
```

```
  A      B      X      C      O      alpha  stderr  Zscore
  <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>
1 Altai Vindija Europeans Yoruba Chimp 0.0282 0.00411 6.85
```

Accomplishing the same task using shell scripting would require a significant amount of work.

Bibliography

1. W. King, The reputed fossil man of the Neanderthal. *Q. J. Sci. Lit. Arts* **1**, 88–97 (1864).
2. A. Hrdlička, The Neanderthal Phase of Man. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* **57**, 249–274 (1927).
3. H. Valladas, G. Valladas, O. Bar-Yosef, B. Vandermeersch, Thermoluminescence dating of Neanderthal and early modern humans in the Near East. *Endeavour* **15**, 115–119 (1991).
4. R. Grün, *et al.*, U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J. Hum. Evol.* **49**, 316–334 (2005).
5. J. Krause, *et al.*, Neanderthals in central Asia and Siberia. *Nature* **449**, 902–904 (2007).
6. C. B. Stringer, J. Hublin, New age estimates for the Swanscombe hominid, and their significance for human evolution. *J. Hum. Evol.* **37**, 873–877 (1999).
7. J. L. Bischoff, *et al.*, High-resolution U-series dates from the Sima de los Huesos hominids yields 600–66+ ∞ kyrs: implications for the evolution of the early Neanderthal lineage. *J. Archaeol. Sci.* **34**, 763–770 (2007).
8. T. Higham, *et al.*, The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512**, 306–309 (2014).
9. E. Trinkaus, *et al.*, An early modern human from the Peștera cu Oase, Romania. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11231–11236 (2003).
10. H. Rougier, *et al.*, Peștera cu Oase 2 and the cranial morphology of early modern Europeans. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1165–1170 (2007).
11. S. Benazzi, *et al.*, Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479**, 525–528 (2011).
12. Q. Fu, *et al.*, An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
13. J.-J. Hublin, *et al.*, Initial Upper Palaeolithic Homo sapiens from Bacho Kiro Cave, Bulgaria. *Nature* **581**, 299–302 (2020).
14. R. L. Cann, M. Stoneking, A. C. Wilson, Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
15. C. Duarte, *et al.*, The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 7604–7609 (1999).
16. M. H. Wolpoff, J. Hawks, D. W. Frayer, K. Hunley, Modern human ancestry at the peripheries: a test of the replacement theory. *Science* **291**, 293–297 (2001).
17. D. W. Frayer, J. Jelínek, M. Oliva, M. H. Wolpoff, “Aurignacian Male Crania, Jaws and Teeth from the Mladeč Caves, Moravia, Czech Republic” in *Early Modern Humans at*

- the Moravian Gate: The Mladeč Caves and Their Remains*, M. Teschler-Nicola, Ed. (Springer Vienna, 2006), pp. 185–272.
18. M. H. Wolpoff, D. W. Frayer, J. Jelinek, “Aurignacian Female Crania and Teeth from the Mladeč Caves, Moravia, Czech Republic” in *Early Modern Humans at the Moravian Gate: The Mladeč Caves and Their Remains*, M. Teschler-Nicola, Ed. (Springer Vienna, 2006), pp. 273–340.
 19. A. Soficaru, A. Dobos, E. Trinkaus, Early modern humans from the Peștera Muierii, Baia de Fier, Romania. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17196–17201 (2006).
 20. E. Trinkaus, European early modern humans and the fate of the Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7367–7372 (2007).
 21. A. Soficaru, C. Petrea, A. Doboș, E. Trinkaus, The Human Cranium from the Peștera Cioclovina Uscată, Romania: Context, Age, Taphonomy, Morphology, and Paleopathology. *Curr. Anthropol.* **48**, 611–619 (2007).
 22. G. Bräuer, H. Broeg, C. B. Stringer, “Earliest Upper Paleolithic crania from Mladeč, Czech Republic, and the question of Neanderthal-modern continuity: metrical evidence from the fronto-facial region” in *Neanderthals Revisited: New Approaches and Perspectives*, J.-J. Hublin, K. Harvati, T. Harrison, Eds. (Springer Netherlands, 2006), pp. 269–279.
 23. K. Harvati, P. Gunz, D. Grigorescu, Cioclovina (Romania): affinities of an early modern European. *J. Hum. Evol.* **53**, 732–746 (2007).
 24. S. E. Bailey, T. D. Weaver, J.-J. Hublin, Who made the Aurignacian and other early Upper Paleolithic industries? *J. Hum. Evol.* **57**, 11–26 (2009).
 25. J. J. Hublin, F. Spoor, M. Braun, F. Zonneveld, S. Condemi, A late Neanderthal associated with Upper Palaeolithic artefacts. *Nature* **381**, 224–226 (1996).
 26. R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, A. C. Wilson, DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**, 282–284 (1984).
 27. S. Pääbo, Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**, 644–645 (1985).
 28. K. B. Mullis, F. A. Faloona, “Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction” in *Methods in Enzymology*, (Academic Press, 1987), pp. 335–350.
 29. M. Krings, *et al.*, Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
 30. D. Serre, *et al.*, No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* **2**, E57 (2004).
 31. R. E. Green, *et al.*, A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416–426 (2008).

32. A. W. Briggs, *et al.*, Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009).
33. M. Nordborg, On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63**, 1237–1240 (1998).
34. R. E. Green, *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
35. E. S. Lander, *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
36. J. P. Noonan, *et al.*, Sequencing and analysis of Neandertal genomic DNA. *Science* **314**, 1113–1118 (2006).
37. R. E. Green, *et al.*, Analysis of one million base pairs of Neandertal DNA. *Nature* **444**, 330–336 (2006).
38. K. Prüfer, *et al.*, The complete genome sequence of a Neandertal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
39. K. Prüfer, *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
40. F. Mafessoni, *et al.*, A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15132–15136 (2020).
41. J. Krause, *et al.*, The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894–897 (2010).
42. D. Reich, *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
43. M. Meyer, *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
44. D. Reich, *et al.*, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).
45. F. Chen, *et al.*, A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019).
46. S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).
47. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
48. S. Mallick, *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
49. J. D. Wall, *et al.*, Higher levels of neandertal ancestry in East Asians than in Europeans.

- Genetics* **194**, 199–209 (2013).
50. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
 51. S. Sankararaman, *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
 52. B. Vernot, *et al.*, Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
 53. I. Lazaridis, *et al.*, Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
 54. B. Y. Kim, K. E. Lohmueller, Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. *Am. J. Hum. Genet.* **96**, 454–461 (2015).
 55. B. Vernot, J. M. Akey, Complex history of admixture between modern humans and Neandertals. *Am. J. Hum. Genet.* **96**, 448–453 (2015).
 56. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**, 677–687.e16 (2020).
 57. L. Skov, *et al.*, Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**, e1007641 (2018).
 58. J. K. Pickrell, *et al.*, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2632–2637 (2014).
 59. M. Petr, S. Pääbo, J. Kelso, B. Vernot, Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1639–1644 (2019).
 60. Q. Fu, *et al.*, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
 61. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–61.e9 (2018).
 62. F. Racimo, S. Sankararaman, R. Nielsen, E. Huerta-Sánchez, Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
 63. R. Nielsen, *et al.*, Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
 64. M. Dannemann, F. Racimo, Something old, something borrowed: admixture and adaptation in human evolution. *Curr. Opin. Genet. Dev.* **53**, 1–8 (2018).
 65. S. Sankararaman, S. Mallick, N. Patterson, D. Reich, The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* **26**,

- 1241–1247 (2016).
66. E. Huerta-Sánchez, *et al.*, Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
 67. R. M. Gitterman, *et al.*, Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Curr. Biol.* **26**, 3375–3382 (2016).
 68. M. Dannemann, A. M. Andrés, J. Kelso, Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am. J. Hum. Genet.* **98**, 22–33 (2016).
 69. D. Enard, D. A. Petrov, Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **175**, 360–371.e13 (2018).
 70. Q. Fu, *et al.*, The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
 71. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
 72. J. P. Bollback, T. L. York, R. Nielsen, Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).
 73. Q. Fu, *et al.*, DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2223–2227 (2013).
 74. A. Seguin-Orlando, *et al.*, Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113–1118 (2014).
 75. K. Harris, R. Nielsen, The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881–891 (2016).
 76. I. Juric, S. Aeschbacher, G. Coop, The Strength of Selection against Neanderthal Introgression. *PLoS Genet.* **12**, e1006340 (2016).
 77. K. Harris, R. Nielsen, Q&A: Where did the Neanderthals go? *BMC Biol.* **15**, 73 (2017).
 78. C. N. Simonti, *et al.*, The phenotypic legacy of admixture between modern humans and Neanderthals. *Science* **351**, 737–741 (2016).
 79. R. C. McCoy, J. Wakefield, J. M. Akey, Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* **168**, 916–927.e12 (2017).
 80. M. Dannemann, K. Prüfer, J. Kelso, Functional implications of Neanderthal introgression in modern humans. *Genome Biol.* **18**, 61 (2017).
 81. M. Silvert, L. Quintana-Murci, M. Rotival, Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am. J. Hum. Genet.* **104**, 1241–1250 (2019).
 82. N. Telis, R. Aguilar, K. Harris, Selection against archaic hominin genetic variation in

- regulatory regions. *Nat Ecol Evol* **4**, 1558–1566 (2020).
83. M. Meyer, *et al.*, A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403–406 (2014).
 84. M. Meyer, *et al.*, Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
 85. C. Posth, *et al.*, Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat. Commun.* **8**, 16046 (2017).
 86. M. Kuhlwilm, *et al.*, Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
 87. M. J. Hubisz, A. L. Williams, A. Siepel, Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* **16**, e1008895 (2020).
 88. S. Peyrégne, *et al.*, Nuclear DNA from two early Neanderthals reveals 80,000 years of genetic continuity in Europe. *Sci Adv* **5**, eaaw5873 (2019).
 89. F. L. Mendez, G. D. Poznik, S. Castellano, C. D. Bustamante, The Divergence of Neanderthal and Modern Human Y Chromosomes. *Am. J. Hum. Genet.* **98**, 728–734 (2016).
 90. C. Glenn Begley, L. M. Ellis, Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
 91. Open Science Collaboration, PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
 92. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature*, 452 (2016).
 93. D. Fanelli, Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631 (2018).
 94. G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9**, e1003285 (2013).
 95. J. A. Papin, F. Mac Gabhann, H. M. Sauro, D. Nickerson, A. Rampadarath, Improving reproducibility in computational biology research. *PLoS Comput. Biol.* **16**, e1007881 (2020).
 96. T. Günther, *et al.*, Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11917–11922 (2015).
 97. M. Raghavan, *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
 98. M. Gallego Llorente, *et al.*, Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015).

99. I. Mathieson, *et al.*, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
100. Z. Hofmanová, *et al.*, Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6886–6891 (2016).
101. C. M. Schlebusch, *et al.*, Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
102. M. Hajdinjak, *et al.*, Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652–656 (2018).
103. V. M. Narasimhan, *et al.*, The formation of human populations in South and Central Asia. *Science* **365** (2019).
104. Y. Xue, *et al.*, Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242–245 (2015).
105. L. A. F. Frantz, *et al.*, Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science* **352**, 1228–1231 (2016).
106. J. I. Meier, *et al.*, Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* **8**, 14363 (2017).
107. I. Lazaridis, *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
108. I. Olalde, *et al.*, The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
109. M. Petr, B. Vernot, J. Kelso, admixr-R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* **35**, 3194–3195 (2019).
110. R Core Team, R: A Language and Environment for Statistical Computing (2020).
111. R. F. Nespolo, *et al.*, An Out-of-Patagonia migration explains the worldwide diversity and distribution of *Saccharomyces eubayanus* lineages. *PLoS Genet.* **16**, e1008777 (2020).
112. G. Debortoli, *et al.*, Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data. *Sci. Rep.* **10**, 10075 (2020).
113. K. Matsudaira, T. Ishida, Divergence and introgression in small apes, the genus *Hylobates*, revealed by reduced representation sequencing. *Cold Spring Harbor Laboratory*, 2020.05.31.126078 (2020).
114. S. Surbakti, *et al.*, New Guinea highland wild dogs are the original New Guinea singing dogs. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24369–24376 (2020).
115. J. Mulindwa, *et al.*, High Levels of Genetic Diversity within Nilo-Saharan Populations: Implications for Human Adaptation. *Am. J. Hum. Genet.* **107**, 473–486

- (2020).
116. D. Liu, *et al.*, Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. *Mol. Biol. Evol.* **37**, 2503–2519 (2020).
 117. M. Ravinet, M. Kume, A. Ishikawa, J. Kitano, Patterns of genomic divergence and introgression between Japanese stickleback species with overlapping breeding habitats. *J. Evol. Biol.* (2020) <https://doi.org/10.1111/jeb.13664>.
 118. Y. Y. Yamasaki, *et al.*, Genome-wide patterns of divergence and introgression after secondary contact between *Pungitius* sticklebacks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190548 (2020).
 119. R. S. Taylor, *et al.*, The role of introgression and ecotypic parallelism in delineating intraspecific conservation units. *Mol. Ecol.* **29**, 2793–2809 (2020).
 120. C. B. Cole, S. J. Zhu, I. Mathieson, K. Prüfer, G. Lunter, Ancient Admixture into Africa from the ancestors of non-Africans. *Cold Spring Harbor Laboratory*, 2020.06.01.127555 (2020).
 121. O. François, F. Jay, Factor analysis of ancient population genomic samples. *Nat. Commun.* **11**, 4661 (2020).
 122. B. Dobon, *et al.*, The shaping of immunological responses through natural selection after the Roma Diaspora. *Sci. Rep.* **10**, 16134 (2020).
 123. J. Kelleher, A. M. Etheridge, G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
 124. D. Massilani, *et al.*, Denisovan ancestry and population history of early East Asians. *Science* **370**, 579–583 (2020).
 125. W. Liu, *et al.*, The earliest unequivocally modern humans in southern China. *Nature* **526**, 696–699 (2015).
 126. K. Harvati, *et al.*, Apidima Cave fossils provide earliest evidence of *Homo sapiens* in Eurasia. *Nature* **571**, 500–504 (2019).
 127. V. Slon, *et al.*, The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018).
 128. S. Colaco, D. Modi, Genetics of the human Y chromosome and its association with male infertility. *Reprod. Biol. Endocrinol.* **16**, 14 (2018).

Dedication

Acknowledgements

Curriculum Vitae

Declaration of Independent Work

I hereby declare that I have conceived and written this thesis independently, without any inadmissible help, and without using resources other than those explicitly stated. I have not previously attempted to complete this or any other doctorate degree.

Leipzig, November 5th, 2020

Martin Petr

Author Contribution Statements

Author contribution statement, Martin Petr
Natural selection and demography in ancient human introgression

Author contribution statement

Title: Limits of long-term selection against Neandertal introgression

Journal: Proceedings of the National Academy of Sciences of the United States of America (PNAS)

Authors: Martin Petr, Svante Pääbo, Janet Kelso and Benjamin Vernot

Part Martin Petr (first author):

- quality control and analysis of published sequencing data
- programmed simulations of demographic history and negative selection
- performed statistical analyses
- wrote the manuscript and supplementary materials

Part Svante Pääbo:

- supervised the project
- wrote the manuscript

Part Janet Kelso (senior author):

- supervised the project
- wrote the manuscript and supplementary materials

Part Benjamin Vernot (senior author):

- contributed statistical analyses
- supervised the project
- wrote the manuscript and supplementary materials

Martin Petr

Janet Kelso

Benjamin Vernot

Author contribution statement, Martin Petr
Natural selection and demography in ancient human introgression

Author contribution statement

Title: The evolutionary history of Neanderthal and Denisovan Y chromosomes

Journal: Science

Authors: Martin Petr, Mateja Hajdinjak, Qiaomei Fu, Elena Essel, H el ene Rougier, Isabelle Crevecoeur, Patrick Semal, Liubov V. Golovanova, Vladimir B. Doronichev, Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Michael V. Shunkov, Maxim B. Kozlikin, Anatoli P. Derevianko, Benjamin Vernot, Matthias Meyer, Janet Kelso

Part Martin Petr (first author):

- quality control and analysis of the new and previously-published sequencing data
- developed statistics to estimate split times
- programmed simulations of selection on introgressed Y chromosomes
- performed statistical analyses
- wrote the manuscript and supplementary materials

Part Mateja Hajdinjak:

- performed ancient DNA sampling
- generated DNA libraries
- wrote supplementary materials

Part Qiaomei Fu:

- designed DNA capture probes
- generated DNA libraries

Part Elena Essel:

- performed ancient DNA sampling
- generated DNA libraries

Part H el ene Rougier:

- provided archaeological samples

Part Isabelle Crevecoeur:

- provided archaeological samples

Part Patrick Semal:

- provided archaeological samples

Part Liubov V. Golovanova:

- provided archaeological samples

Part Vladimir B. Doronichev:

- provided archaeological samples

Part Carles Lalueza-Fox:

- provided archaeological samples

Part Marco de la Rasilla:

- provided archaeological samples

Part Antonio Rosas:

- provided archaeological samples

Part Michael V. Shunkov:

- provided archaeological samples

Part Maxim B. Kozlikin:

- provided archaeological samples

Part Anatoli P. Derevianko:

- provided archaeological samples

Part Benjamin Vernot:

- contributed to statistical analyses
- supervised the project

Part Matthias Meyer:

- supervised the project
- wrote the manuscript and supplementary materials

Part Janet Kelso (senior author):

- supervised the project
- wrote the manuscript and supplementary materials

Martin Petr

Janet Kelso

Author contribution statement, Martin Petr
Natural selection and demography in ancient human introgression

Author contribution statement

Title: *admixr*—R package for reproducible analyses using ADMIXTOOLS

Journal: Bioinformatics

Authors: Martin Petr, Benjamin Vernot, Janet Kelso

Part Martin Petr (first author):

- programmed the *admixr* software package
- wrote the manuscript and supplementary materials

Part Benjamin Vernot:

- contributed to software testing

Part Janet Kelso (senior author):

- supervised the project
- wrote the manuscript and supplementary materials

Martin Petr

Janet Kelso