

# The evolutionary history of human spindle genes includes back-and-forth gene flow with Neandertals

**Authors:** Stéphane Peyrégne\*, Janet Kelso, Benjamin M. Peter, Svante Pääbo\*

**Affiliation:** Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany.

**\*Correspondence to:** stephane\_peyregne@eva.mpg.de, paabo@eva.mpg.de

**Author Contributions:** S.Pe. and S.Pä. designed research; S.Pe. performed research; S.Pe., J.K., B.M.P. and S. Pä. interpreted the results; S.Pe. and S.Pä. wrote the article with input from all authors.

**Competing Interest Statement:** The authors declare no competing interest.

**Key words:** Human evolution, Neandertals, gene flow, positive selection, spindle

# Abstract

Proteins associated with the spindle apparatus, a cytoskeletal structure that ensures the proper segregation of chromosomes during cell division, experienced an unusual number of amino acid substitutions in modern humans after the split from the ancestors of Neandertals and Denisovans. Here, we analyze the history of these substitutions and show that some of the genes in which they occur may have been targets of positive selection. We also find that the two changes in the kinetochore scaffold 1 (KNL1) protein, previously believed to be specific to modern humans, were present in some Neandertals. We show that the *KNL1* gene of these Neandertals shared a common ancestor with present-day Africans about 200,000 years ago due to gene flow from the ancestors (or relatives) of modern humans into Neandertals. Subsequently, some non-Africans inherited this modern human-like gene variant from Neandertals, but none inherited the ancestral gene variants. These results add to the growing evidence of early contacts between modern humans and archaic groups in Eurasia and illustrate the intricate relationships among these groups.

# Introduction

The ancestors of Neandertals and Denisovans diverged from those of modern humans between 522 and 634 thousand years ago (kya) (1). Differences between modern humans and Neandertals, Denisovans and other hominins can therefore reveal traits that changed in modern humans over the past half-million years and are specific to modern humans. The paleontological and archaeological records provide information about morphological and cultural traits (2, 3), yet other traits remain inaccessible by such approaches. The retrieval of DNA from archaic human remains and the reconstruction of Neandertal and Denisovan genomes (4-6) open an additional approach to understanding what sets modern and archaic humans apart.

Because large numbers of present-day human genomes have been sequenced while few archaic genomes are available, the comparisons of Neandertal and Denisovan genomes are particularly useful for identifying those nucleotide changes that occurred on the modern human lineage and reached fixation (or high frequencies) in present-day people. According to one estimate (6), there are 31,389 such differences of which only 96 change the amino acid sequence of 87 proteins. Three of these missense changes occur in a single gene, *sperm associated antigen 5* (*SPAG5*), which encodes a protein associated with the spindle apparatus (7, 8). The spindle is a cytoskeletal structure composed of microtubules and associated proteins that attach to the chromosomes and ensures their proper segregation during cell division (9). *SPAG5* is the only gene in the genome with three missense changes, but four of the 87 genes carry two missense changes. One of these is *kinetochore scaffold 1* (*KNL1*, previously *CASC5*) which encodes a protein in the kinetochore (10), a protein structure at the centromere of chromosomes to which the spindle attaches during mitosis and meiosis. The occurrence of three missense changes in *SPAG5* and two in *KNL1*, as well as one in the gene *KIF18A*, which encodes a protein involved in the movement of chromosomes along microtubules (11), is intriguing as it suggests

that components of the spindle may have been subject to natural selection during recent human evolution (12, 13).

We therefore set out to study genes with missense changes that are associated with the spindle apparatus and to reconstruct their evolutionary history in modern and archaic humans. We show that multiple spindle genes may have experienced positive selection on the modern human lineage. We also show that the variant of *KNL1* that carries two derived missense mutations was transferred by gene flow from modern humans to early Neandertals and then again from late Neandertals to modern humans.

## Results

### Spindle-associated genes with missense changes in modern humans.

According to the classification of genes in the Gene Ontology database (14, 15) that represents current knowledge about the function of genes, there are eight spindle-associated genes among the 87 genes that carry missense changes that are fixed or almost fixed among present-day humans. These eight genes are *SPAG5* and *KNL1*, with three and two changes, respectively, and *ATRX*, *KATNA1*, *KIF18A*, *NEK6*, *RSPH1* and *STARD9*, which each carry one missense change. A total of eleven missense changes accumulated in these eight genes. Given the length of spindle-associated genes, six missense changes are expected making this a significant enrichment of amino acid changes compared to random expectation (permutation test  $p = 0.045$ ; see Methods). Multiple missense changes in a single gene are also more than expected by chance ( $p = 0.044$  for two changes;  $p < 0.001$  for three changes). The genes with single changes have a range of functions associated with cell division. *ATRX* is a chromatin remodeler required for normal chromosome alignment, cohesion and segregation during meiosis and mitosis (16, 17). *KATNA1* and *KIF18A* are depolymerases that disassemble microtubules in the spindle apparatus and are essential for proper chromosome alignment at the equator of dividing cells (11, 18). *NEK6* is a kinase that is required for cell cycle progression through mitosis (19) and interacts with proteins of the centrosome (20), an organelle that serves as a microtubule-organizing center and is crucial for the spindle apparatus. *RSPH1* localizes to the spindle and chromosomes during meiotic metaphase (21), when the chromosomes align in the dividing cell. Finally, *STARD9* localizes to the centrosome during cell division and is required for spindle assembly by maintaining centrosome integrity (22). We describe the location and predicted effects of the missense changes in Appendix 1 – Tables 1 and 2.

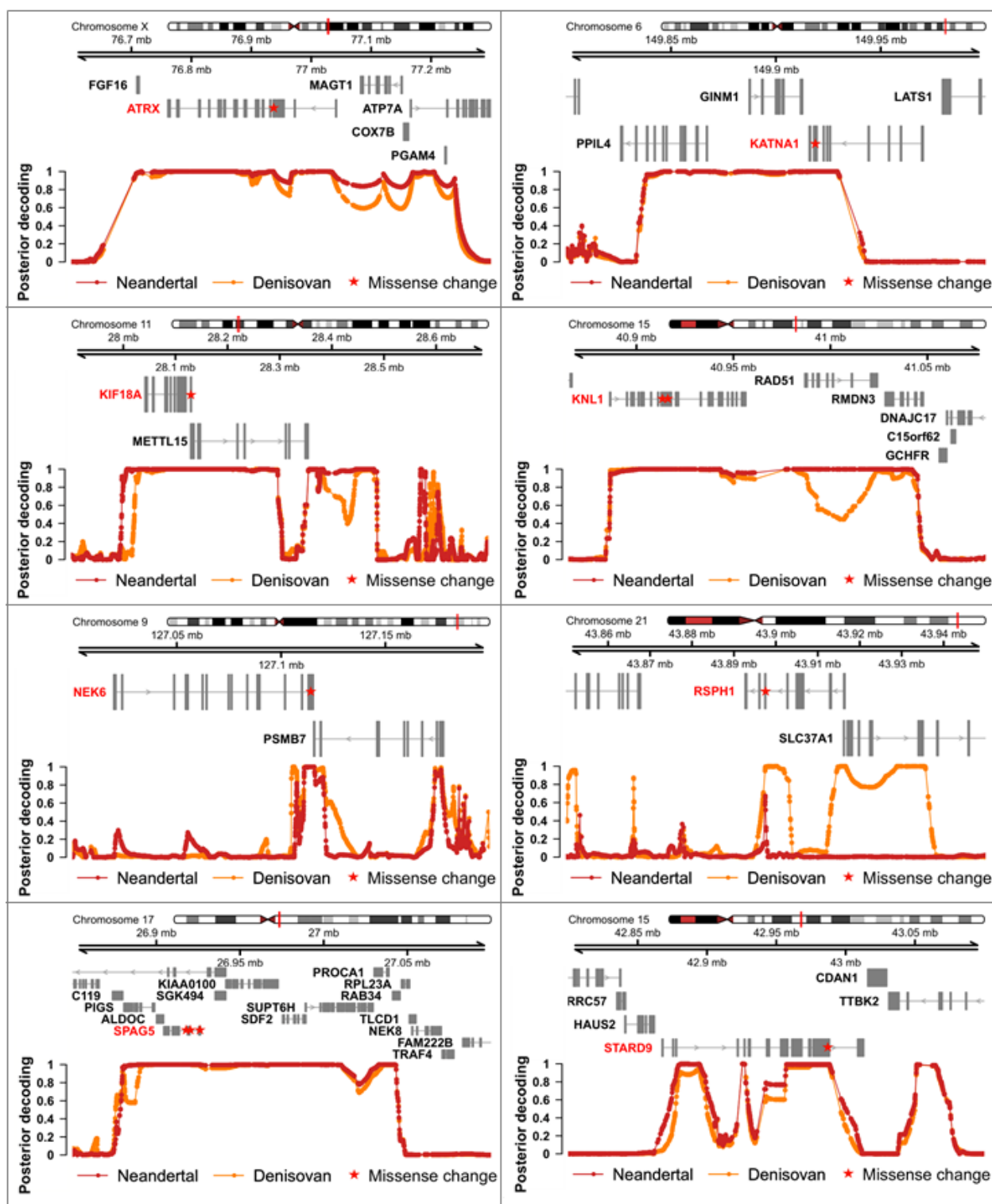
### Evolutionary history of spindle proteins in modern humans.

The enrichment of missense changes in genes associated with the spindle apparatus suggests that some of these changes may have been subject to positive selection in modern humans (23). To test whether the eleven changes in the eight genes may have been beneficial, we applied a method to detect positive selection that occurred in the common ancestral population of modern humans after their split from archaic humans (24). This method relies on a hidden Markov model to identify genomic segments where archaic human genomes fall outside of modern human variation, *i.e.*, where all modern humans share a common ancestor more recently than any common ancestor shared with archaic humans. Applying this

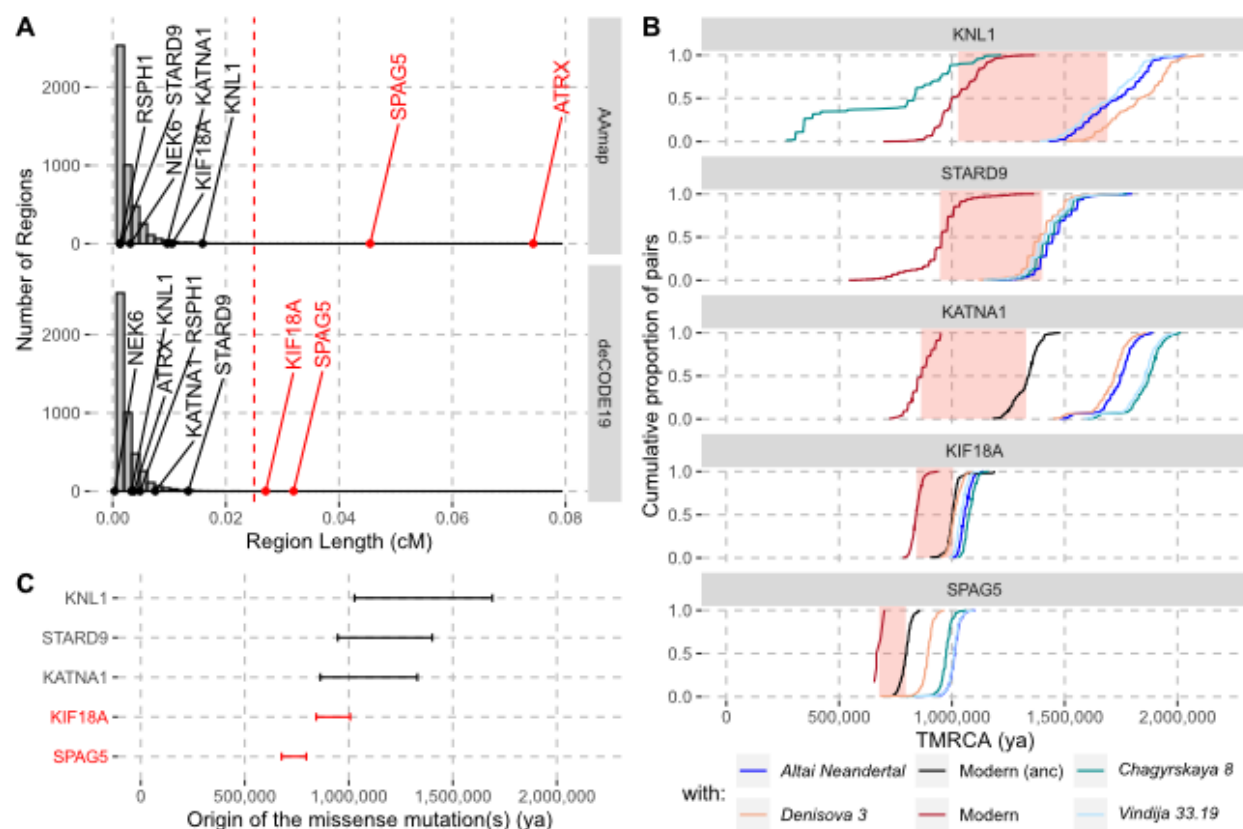
approach to the genomes of a Neandertal (6), a Denisovan (5) and 504 Africans (25), we identified such segments spanning 132 to 547,448 base pairs (bp) around the missense substitutions of each of the eight spindle genes (Figure 1; Supplement File 1). The genetic lengths of these segments are informative about how fast they spread in modern humans since the common ancestor shared with archaic humans. Segments longer than 0.025cM were not observed in neutral simulations (i.e., false positive rate <0.1%, (24)) and therefore represent evidence for positive selection. We account for uncertainty in local recombination rates by using two recombination maps (African-American and deCODE maps, (26, 27)). One gene, *SPAG5*, carries a segment around the missense substitutions longer than 0.025cM in both maps (Figure 2). The segments in two genes, *ATRX* and *KIF18A*, exceed this cutoff only for the African-American and deCODE recombination maps, respectively (Figure 2A). Thus, there is consistent evidence for positive selection on *SPAG5*, while evidence for positive selection affecting *ATRX* and *KIF18A* is dependent on the recombination maps used.

To gain further insights into the history of the missense substitutions in the spindle genes, we estimated the time when each of these mutations occurred. Note that this time differs from that of fixation, which may have happened much later, particularly in the absence of positive selection. As fixation events since the split with the ancestors of Neandertals and Denisovans are rare (24), it is most parsimonious to assume that the mutations that reached fixation in the region of each spindle gene represent a single event of fixation (i.e., were linked to the missense substitution(s) as they spread in modern humans). Thus, estimates of the coalescence time of the haplotypes in which the substitutions are found today provides a most recent bound for when the mutations occurred. By contrast, the most recent common ancestors shared between modern and archaic humans (who carried the ancestral variants) will predate the occurrence of the mutations and therefore provide older bounds for the time of origin of the missense variants.

By computing pairwise differences among the high-quality genome sequences of 104 individuals from Africa (Human Genome Diversity Panel, (28); including two non-African individuals with the ancestral version of *KATNA1* and *KIF18A*, respectively; Methods), and four archaic human genomes (1, 5, 6, 29), we estimated the ages of the missense substitutions of *KATNA1*, *KIF18A*, *KNL1*, *SPAG5* and *STARD9* (Figure 2B and 2C, Appendix 2 - Table 1). We excluded *NEK6* and *RSPH1* as the regions identified around the missense substitutions were too short to estimate their age (4,104bp and 132bp, respectively). We also excluded *ATRX*, which is located on the X chromosome. The relative difference in age estimates suggests that the mutations in *KATNA1*, *KNL1* and *STARD9* occurred much earlier than the mutations in *SPAG5* and *KIF18A* and may be more than a million years old. By contrast, the mutations in *SPAG5* and *KIF18A* are more recent. This is consistent with that the latter two genes have some evidence of positive selection around their missense substitutions whereas the former genes lack evidence for selection.



**Figure 1: Genomic regions around spindle genes where archaic humans fall outside the modern human variation.** Each panel corresponds to the region around the missense change(s) (red stars) in a spindle gene. The grey boxes correspond to exons. The curves give the posterior probability (computed as in (24)) that an archaic genome (*Altai Neandertal* in red, *Denisova 3* in orange) is an outgroup to present-day African genomes at a particular position (dots on the curves correspond to informative positions, i.e., polymorphic positions or fixed derived substitutions in Africans from the 1000 Genomes Project phase III, compared to four ape genomes). Chromosomal locations are given on top.



**Figure 2: Evidence for selection in the spindle genes with age estimates of these substitutions.** (A) The genetic length of segments around the missense substitutions where the *Altai Neandertal* and *Denisova 3* fall outside the human variation (Figure 1) using the African-American map, Amap, or the deCODE map, deCODE19. The grey histogram corresponds to the length distribution of such segments in neutral simulations (24). Candidate genes for selection (red) are those with segments longer than 0.025cM (24) (vertical red dashed line). (B) Cumulative distributions of pairwise times to the most recent common ancestors (TMRCA) among present-day African chromosomes with the most distant relationships (red; see Methods), or between the chromosomes of present-day Africans and either present-day individuals with the ancestral versions of the missense substitutions ("Modern (anc)", in black) or archaic humans (other colors). The pink areas correspond to estimated time intervals for the origin of the missense substitutions and their bounds correspond to the average TMRCA over the red curve and the next one (back in time), respectively. (C) Summary of ages of substitutions as described in panel B. Genes with evidence of positive selection are highlighted in red.

## A modern human-like *KNL1* haplotype in Neandertals.

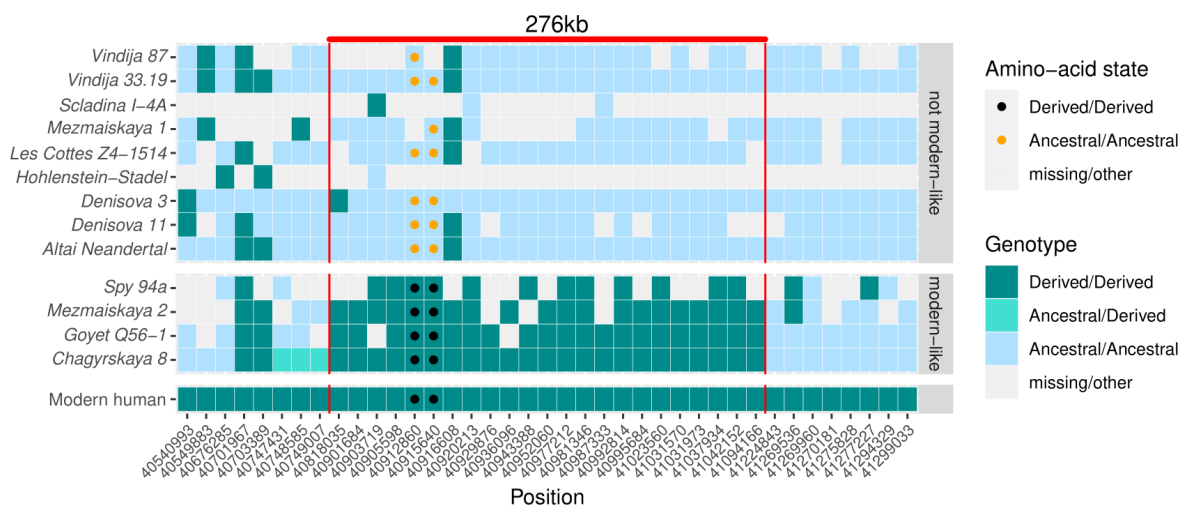
The identification of modern human-specific missense changes was based on the high-coverage genomes of one Neandertal and one Denisovan (6). With the availability of additional archaic human genome sequences, it is now possible to explore whether the derived states may have occurred in some archaic humans. For five of the spindle genes, sequence information from seven to ten archaic humans is available at the relevant positions (1, 6, 29-32). In none of them, there is evidence for the presence of the derived missense variants (Appendix 3 – Table 1). For one gene, *STARD9*, one DNA fragment sequenced from a Neandertal individual (*Denisova 5*, 52-fold average sequence coverage) carries the derived variant whereas 29 carry the ancestral variant. This is likely to represent present-day human DNA contamination or a sequencing error. For another gene, *SPAG5*, which carries three missense substitutions, two DNA fragments sequenced from a Neandertal individual (*Mezmaiskaya 1*, 1.9-fold

average sequence coverage) carry the derived variant at one of these positions (chr17:26,919,034; hg19) whereas three DNA fragments carry the ancestral variant. No information was available for this individual at the two other positions, but none of eight neighbouring positions where present-day Africans carry fixed derived alleles and where information is available for this Neandertal carries any modern human-like allele (Appendix 3 – Table 2). This could therefore represent present-day human DNA contamination, which amounts to 2-3% in the DNA libraries sequenced from this individual (1).

In contrast, between one and 27 DNA fragments that cover the two missense substitutions in *KNL1* in four of the twelve Neandertals available carried only derived alleles (Figure 3). This includes the *Chagyrskaya 8* genome, which is sequenced to 27-fold average genomic coverage. Several of the fragments carrying derived alleles also carry cytosine-to-thymine substitutions near their ends, which is typical of ancient DNA molecules and suggests that the fragments do not represent present-day human DNA contamination (33).

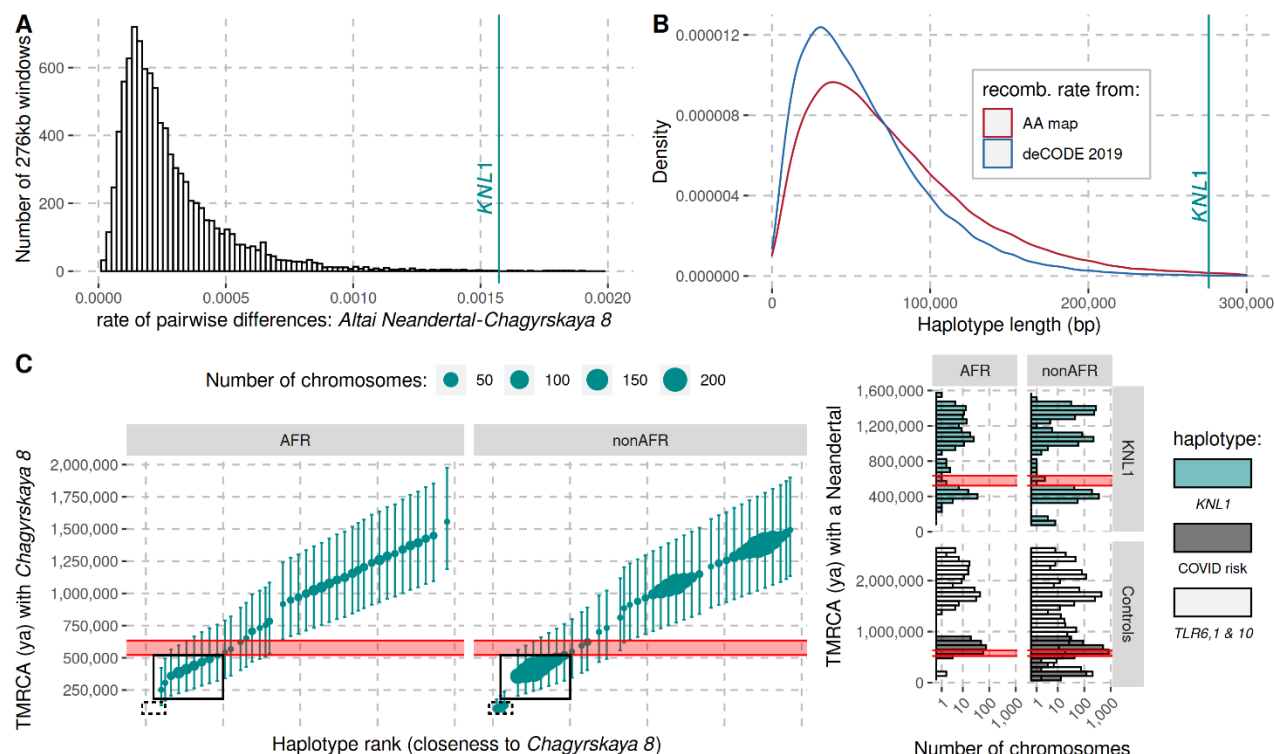
Further evidence that the modern human-like alleles in *KNL1* in the four Neandertals are not due to present-day DNA contamination comes from the observation that they carry modern human-like derived alleles that occur in all present-day Africans at 21 other positions in a 276kb-long region around the missense variants, whereas the other Neandertal and Denisovan individuals who carry the ancestral alleles do not (Figure 3). In addition, as there is no evidence of ancestral alleles at any of these informative positions, the four Neandertals probably all carried this “modern human-like” haplotype in homozygous form.

To investigate how the divergence among Neandertals of the haplotypes with and without the derived missense substitutions in the *KNL1* gene compares to other regions of Neandertal genomes, we calculated the divergence in non-overlapping 276kb-segments between the Altai Neandertal (*Denisova 5*), who carries ancestral *KNL1* substitutions, and *Chagyrskaya 8*, who carries the derived substitutions (Figure 4A). The number of differences in the *KNL1* region is about one order of magnitude higher than the average across the genome and in the top 0.27% of all regions in the genome. Thus, the *KNL1* region stands out as unusually diverged between these two Neandertals.



**Figure 3: A modern human-like haplotype in some Neandertals.** Genotypes from 13 archaic individuals (y-axis) are shown in a region around the two missense changes (dots) in *KNL1*. We only show positions (x-axis) that are derived in all Luhya and Yoruba individuals from the 1000 Genomes Project compared to four great apes (24) and at least one high-coverage archaic

genome (*Chagyrskaya 8*, *Denisova 3*, *Vindija 33.19* and *Altai Neandertal*, i.e., *Denisova 5*). The colors of the squares and dots represent the genotype, with ancestral and derived alleles. For the low coverage archaic genomes, we randomly sampled a sequence at each position. Red lines indicate the modern human-like haplotype.



**Figure 4: The modern human-like *KNL1* haplotype in Neandertals.** (A) Pairwise differences between two high coverage Neandertal genomes (*Chagyrskaya 8* and *Altai Neandertal (Denisova 5)*) in non-overlapping sliding windows of 276kb (histogram) and in the *KNL1* region (vertical cyan line; chr15:40,818,035-41,094,166, hg19). Windows with less than 10,000 genotype calls for both Neandertals were discarded. (B) The expected length distributions under a model of incomplete lineage sorting based on local recombination rate estimates from the African-American (AA) and deCODE recombination maps and the length of the modern human-like *KNL1* haplotype in Neandertals (vertical cyan line). (C) Left panel: Time to the most recent common ancestor (TMRCA) between the *Chagyrskaya 8* Neandertal (who carries the modern human-like haplotype) and *KNL1* haplotypes in present-day humans with their 95% confidence intervals (bars) for chr15:40,885,107-40,963,160 (hg19). The size of the points corresponds to the number of chromosomes carrying this haplotype in the HGDP dataset. The black rectangles highlight subsets of haplotypes with TMRCAs more recent than the modern-archaic population split time (1) (shaded pink area). Right panel: Distribution of pairwise TMRCAs between the Neandertal and present-day humans from the HGDP dataset in the region of *KNL1* and two other regions with archaic haplotypes in present-day humans (Controls, (34, 35); COVID risk region: chr3:45,859,651-45,909,024; TLR6, 1 & 10: chr4:38,760,338-38,846,385). We used the *Vindija 33.19* genome for the COVID risk haplotype and the *Chagyrskaya 8* genome otherwise.

## Introgression of the *KNL1* haplotype into Neandertals.

It is intriguing that the Neandertals who carry the two missense changes in *KNL1* also carry modern human-like alleles shared by all (or nearly all) present-day Africans at multiple positions in the region of *KNL1* and exhibit unusually high divergence to other Neandertal haplotypes. This raises the question if Neandertals may have inherited this haplotype from ancestors or relatives of modern humans. However, the age of the *KNL1* missense mutations predates the divergence of Neandertals and modern

humans (Figure 2C) and must thus have been segregating in the ancestral populations of the two groups. This may have resulted in the presence of the derived variants in both Neandertals and modern humans even in the absence of any gene flow (“incomplete lineage sorting”). However, in this case, the segment carrying similarity between Neandertal and modern human genomes is expected to be shorter than if it entered the Neandertal population by gene flow because the number of generations over which recombination would have had the opportunity to shorten the haplotype would be larger.

We used local estimates of the recombination rate (Methods) and a published method (36) to infer the expected length distribution for haplotypes inherited from the population ancestral to Neandertals and modern humans. The 276kb haplotype is longer than expected (Figure 4B,  $p \leq 0.006$ ) and it is therefore unlikely to be inherited from the common ancestral population. Thus, although the missense variants were segregating in the common ancestral population of Neandertals and modern humans, Neandertals did not inherit this haplotype from that population. Rather, the haplotype in Neandertals is likely the result of gene flow between Neandertals and modern humans. Since all present-day humans carry the derived *KNL1* variants whereas only some Neandertals do, and since the modern human-like *KNL1* haplotype is unusually diverged to other Neandertal haplotypes, gene flow was likely from modern humans (or their relatives) to Neandertals.

### Age of *KNL1* gene flow into Neandertals.

Using the length of the haplotype, the local recombination rate (averaged over the African-American and deCODE recombination maps), the age of the most recent Neandertal carrying the haplotype (~40kya (30)), and given the limitation that recombination among the fixed alleles in modern humans cannot be detected, we estimate an older age limit for the haplotypes seen in Neandertals and modern humans of 265kya (Methods). This means that gene flow occurred after this time.

It is possible to further refine the age estimate of this gene flow if some present-day humans are closer to the Neandertals carrying the haplotype than other present-day humans. By comparing the high-quality genomes of Neandertals with and without the modern-like *KNL1* haplotype, we identified 206 positions that are derived on the modern-like *KNL1* haplotype in Neandertals. We then computed the frequency of these alleles among 104 present-day African genomes and identified 19 positions where 32-39% of present-day people share the derived allele (Appendix 4 - Table 1). These derived alleles are physically linked ( $r^2 > 0.58$ ) and tag a 78kb-long haplotype (102kb in some individuals,  $r^2 > 0.39$ ) where some present-day individuals are closer to the *Chagyrskaya 8* Neandertal than other present-day individuals. The individuals with the 102kb-long haplotype carry seven differences to the *Chagyrskaya 8* Neandertal (among the 36,106 bases called). This number of differences yields an estimate for the most recent common ancestor with Neandertals carrying the modern-like *KNL1* haplotype of 251kya (95% CI: 131-421kya; Figure 4C). Note that this haplotype is also too long to be inherited from the common ancestral population of Neandertals and modern humans half a million years ago (95% CI for the age of the last common ancestor: 49-304kya). The length of the 102kb-long haplotype in combination with the number of differences to *Chagyrskaya 8* yields an age estimate of 202kya (95% CI: 118-317kya, Methods), consistent with the estimate of gene flow <265kya based on the haplotype length in Neandertals.

In conclusion, some Neandertals carry a haplotype encompassing the *KNL1* gene that carries derived alleles also seen in all present-day humans. The length of this haplotype suggests that it entered

the Neandertal population less than 265kya. Some present-day people carry a haplotype in this region that share a common ancestor with the modern human-like haplotype in Neandertals 202kya. Both estimates represent times after which the contact that contributed this haplotype to Neandertals must have occurred.

# **No evidence for selection on the *KNL1* haplotype in Neandertals.**

The oldest Neandertal carrying the modern human-like *KNL1* haplotype for whom we have genome sequence data is *Chagyrskaya 8*, who lived 60-80kya in the Altai Mountains (29). It is not present in two older Neandertals from Europe (31) and one from Siberia (6), nor in two other archaic individuals who lived around 60-80kya in Siberia and the Caucasus (1, 32). However, it is present in three out of five Neandertals who lived 40-50kya (1, 30) in Europe, which may hint at an increase in its frequency over time.

To test this hypothesis, we investigated how often frequencies of other variants in the genomes of these Neandertals similarly increased over time. We identified 7,881 polymorphic transversions at least 50kb apart where the derived allele was absent among the three early Neandertals but present at least once in the Neandertals who lived 60-80kya and 40-50kya. Among these positions, at least three late Neandertals carried the derived alleles at 2,773 positions (35.2%), suggesting that such changes in frequency were common. Although this analysis is limited by the few Neandertal genomes available, it yields no evidence for positive selection on the *KNL1* variants in Neandertals.

# **Reintroduction of the modern-like *KNL1* haplotype into non-Africans.**

As several late Neandertals carried a modern human-like *KNL1* haplotype, it is possible that it was reintroduced into modern humans when they met and mixed outside Africa approximately 44-54kya (37). We would then expect that some non-Africans would carry a haplotype that is more closely related to that of *Chagyrskaya 8* than the present-day humans analyzed above (those highlighted by a solid box in Figure 4C). Among 825 non-African individuals from around the world, we identified 12 individuals from several populations that carry one chromosome that differs at just 7 to 13 positions from the *Chagyrskaya 8* genome in the *KNL1* region (out of ~140kb with genotype calls in the 276kb region). By comparison, other non-African individuals carry 54 to 179 differences to the *Chagyrskaya 8* genome in this region (Figure 4C). We estimated that the 12 individuals share a most recent common ancestor with *Chagyrskaya 8* about 96 to 139kya in this region of *KNL1* (95% CI: 81-145kya to 95-198kya for the individuals with 7 and 13 differences to *Chagyrskaya 8*, respectively; Methods). These estimates are similar to those computed in Neandertal haplotypes previously described in present-day individuals (Figure 4C, right panel (34, 35)). Adjoining the 3'-end of this *KNL1* region, there are seven positions where these 12 individuals share alleles with archaic genomes (including 4 alleles shared with *Chagyrskaya 8*; Supplement File 2) while no other present-day humans in the dataset do so. These observations suggest that these 12 present-day individuals carry a *KNL1* haplotype inherited from Neandertals.

Although the missense alleles in *KNL1* are fixed among the genomes of 2,504 present-day humans (25), some rare ancestral alleles may exist among present-day human genomes, *e.g.*, due to interbreeding with archaic humans or due to back mutations. We therefore looked at the exome and

whole-genome sequences from the *gnomAD* database (v2.1.1, (38)). There are no ancestral alleles among the exome sequences (out of 227,420 alleles called). One of the 15,684 whole genomes carries one ancestral allele at one of the two positions carrying missense variants. This may be a back mutation (rs755472529; Appendix 5 – Table 1). Thus, although there is evidence that gene flow from Neandertals introduced the derived version of *KNL1* into modern humans, there is no evidence of the ancestral version of *KNL1*, which was also present in Neandertals, in present-day people. This is compatible with that the ancestral variants may have been disadvantageous and therefore eliminated after admixture with Neandertals and Denisovans.

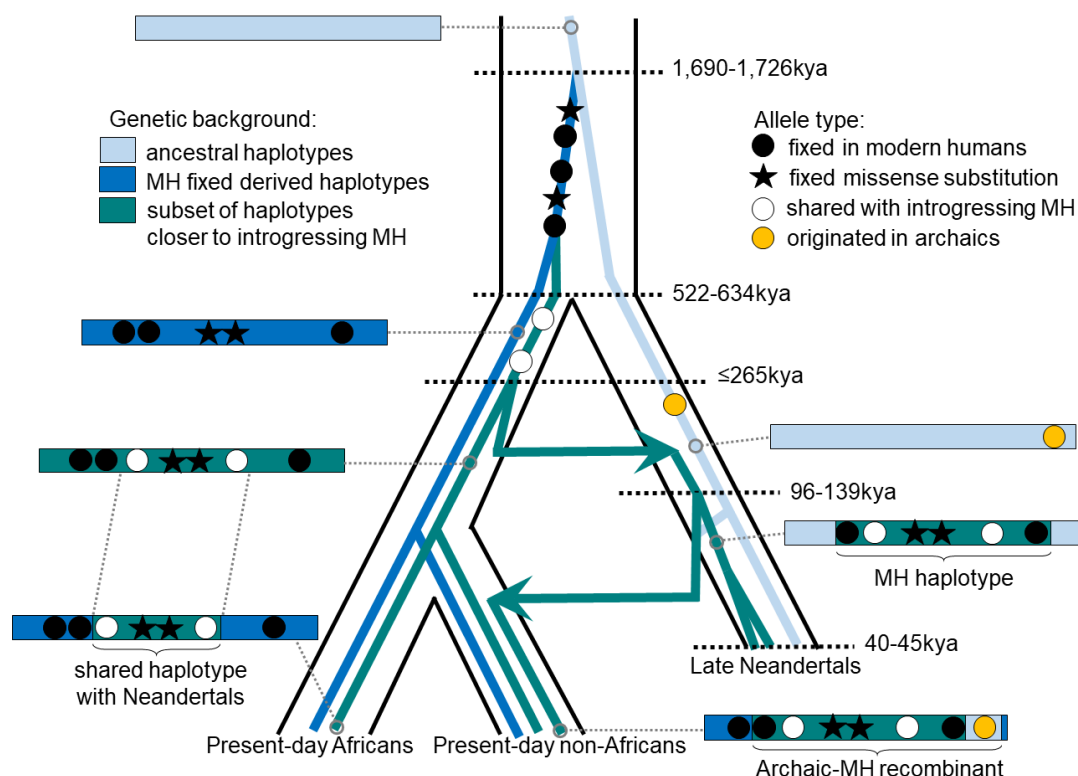
## Discussion

Spindle genes experienced an unusually large number of missense changes in modern humans since the split from a common ancestor with archaic humans. This is intriguing, as mitotic metaphase has been shown to be prolonged in apical progenitors during human brain organoid development when compared to apes (39, 40). Some of the missense changes in spindle genes that occurred since the separation of modern and archaic humans may be involved in such differences.

One of the spindle genes carrying missense changes, *KNL1*, has a particularly interesting evolutionary history in that some Neandertals carried a haplotype sharing two missense variants with present-day humans that were hitherto believed to be specific to modern humans. Both the length of the modern-like *KNL1* haplotype in Neandertals and the small number of differences between present-day humans and Neandertals in this region suggest that they shared a common ancestor more recently than the estimated split time between these two populations. While only a few Neandertals carry this *KNL1* haplotype and its divergence to other Neandertal haplotypes is unusually high (Figure 4A), many variants on the haplotype are fixed or common among present-day humans. Therefore, we infer that ancestors (or relatives) of modern humans contributed this haplotype to Neandertals. Figure 5 summarizes the complex history of *KNL1*.

That groups related to early modern humans contributed a *KNL1* haplotypes to early Neandertals supports previous evidence based on the analyses of mitochondrial DNA (41, 42), Y chromosomes (43) as well as genome-wide data (44, 45) for early contacts between populations related to modern humans and Neandertals in Eurasia. The estimated ages of the most recent common ancestor of the *KNL1* haplotype of ~202kya (based on pairwise differences and length of the shared haplotype) and ≤265kya (based on the length of the haplotype in Neandertals) suggest that at least some contacts were more recent than 265kya and are in agreement with some of the previous age estimates for such contacts (45). This could resolve the apparent discrepancy between the high mitochondrial diversity relative to their nuclear diversity among early Neandertals (31) as the mitochondrial diversity would reflect diversity inherited from a potentially more diverse modern human population. Notably, human remains that may be morphologically similar to modern humans and are of a relevant age have been found in Greece ~210kya (46) and in Israel/Palestine 177–194kya (47). In any event, the complex history of the *KNL1* haplotype illustrates the close and intricate relationship between Neandertals and modern humans.

The enrichment of missense changes in spindle genes (6, 12) suggests that the spindle apparatus may have played a special role during modern human evolution. We show evidence of positive selection in the region around the missense variants in *SPAG5*, and perhaps in *ATRX* and *KIF18A*. Although there is no evidence of selection for the other variants it should be kept in mind that, based on simulations, the method used only detects selective advantages of at least 0.5% with a 65% probability (24). It is also unlikely to detect selection starting from standing variation. Therefore, the absence of evidence for selection does not exclude the possibility of selection. For instance, it is interesting that only derived *KNL1* haplotypes introduced into modern humans from Neandertals persisted until present-day. This provides suggestive evidence that the missense mutations in *KNL1* may also have been important for modern humans. Ultimately, functional characterization of the potential effects of the ancestral and derived variants in spindle genes is necessary to clarify if these changes had effects that could have been advantageous.



**Figure 5: Schematic illustration of the history of *KNL1*.** The tree delineated in black corresponds to the average relationship between the modern and archaic human populations. The inner colored trees correspond to the relationship of different *KNL1* lineages, with arrows highlighting the direction of gene flow between populations. The corresponding haplotypes are illustrated on the sides of the tree and shows the recombination history in the region (e.g., the recombinant Neandertal haplotype with variants of putative archaic origin in non-Africans). Dots correspond to informative positions, and the stars illustrate the missense substitutions. The age of relevant ancestors are marked by horizontal dotted lines. MH: Modern human

## Methods

### Identifying genes associated with the spindle apparatus

The human Gene Ontology Annotation database was downloaded on June 8<sup>th</sup> 2021 (goa\_human.gaf.gz, from <http://current.geneontology.org/products/pages/downloads.html>), together with the basic Gene Ontology terms (<http://purl.obolibrary.org/obo/go/go-basic.obo>). We selected all the Gene Ontology terms that include the keyword “spindle” (65 terms), identified all the human genes that are associated with at least one of these terms (562 genes out of 19,719) and overlapped them with the list of 87 genes that exhibit fixed missense changes in modern humans (6, 12).

### Testing for enrichment of missense changes in spindle genes

The dbNSFP (version 4.2), a database of all potential missense variants in the human genome, was downloaded on October 4<sup>th</sup> 2021 from <ftp://dbnsfp@dbnsfp.softgenetics.com/dbNSFP4.2a.zip> (48). After removing the variants that do not have a coordinate in hg19 and do not pass the minimal filters for the *Altai Neandertal* (*Denisova 5*) and Denisovan (*Denisova 3*) genomes (1), we randomly sampled 96 of the missense variants reported in this database (from the 62,961,368 remaining after filtering) a thousand times and recorded each time how many variants belonged to a spindle gene (as identified in the previous section). If a variant belonged to multiple genes, it was counted only once. The p-value reported in the Results section corresponds to the number of repetitions with eleven or more missense changes in spindle genes. Similarly, to test whether multiple missense changes in a single gene are expected by chance, we counted how often among the thousand repetitions there was at least one gene with two (or three) or more missense changes.

### Processing the human genome datasets and great ape reference genomes

For most analyses, we used phased genotypes from the high-coverage genomes from the Human Genome Diversity Panel (HGDP, (28)), as well as gVCF files for each individual to identify positions which have genotype calls. After subsetting to regions of interest (Appendix 2 – Table 1), gVCF files were converted into VCF files, concatenated with the phased genotypes and merged into a single VCF file using bcftools (49). Genotypes that did not pass all the quality filters from (28) were then discarded. Finally, positions were lifted over to hg19 using picard tools (default parameters, <http://broadinstitute.github.io/picard>) and the chain file hg38ToHg19 from the University of California, Santa Cruz, (UCSC) Genome Browser (<hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz>).

Genotypes from four high-coverage archaic genomes (*Denisova 3* (5), *Altai Neandertal* (6), *Vindija 33.19* (1) and *Chagyrskaya 8* (29)) and nine low-coverage archaic genomes (*Goyet Q56-1*, *Mezmaiskaya 2*, *Les Cottés Z4-1514*, *Vindija 87* and *Spy 94a* from (30), *Hohlenstein-Stadel* and *Scladina I-4A* from (31) and *Mezmaiskaya 1* (1)) were merged into a single VCF file. For the low-coverage archaic genomes, one base with a quality of at least 30 was sampled randomly at each position. Positions outside the mappability track “map35\_100” (i.e., positions outside regions where all overlapping 35-kmers are unique in the genome, (1)) were filtered out, as well as positions without at least one genotype call from the high-coverage archaic genomes. Note that we also checked the bases carried by all sequences overlapping the substitutions in spindle genes from the published BAM files of these archaic human genomes using samtools tview (49).

For each of these datasets, ancestral allele information was retrieved from great ape reference genome assemblies: chimpanzee (panTro4) (50), bonobo (panPan1.1) (51), gorilla (gorGor3) (52), and orangutan (ponAbe2) (53) as LASTZ alignments to the human genome GRCh37/hg19 prepared in-house and by the UCSC Genome Browser (54). We defined the ancestral allele as the one carried by at least three of the four ape reference genome assemblies, allowing a third allele or missing information in only one ape reference genome.

## Processing the recombination maps

To compute genetic distances, we used a recombination map obtained from crossovers between African and European ancestry tracts in African-Americans ((26), available in hg19 coordinates from <http://www.well.ox.ac.uk/~anjali/>) and a map based on crossovers in parent-offspring pairs from Iceland (deCODE, (27)). The latter was lifted over to hg19 (originally in hg38 coordinates) with the liftover tool and the chain file hg38ToHg19 from UCSC (using a minimum match rate of 0.9 between bases of both assemblies). This resulted in both gaps and overlaps between windows of the recombination map. Therefore, we assumed that the recombination rate in a gap was the average of the two directly adjacent windows, and we truncated the windows that overlap with a previous window (or removed the window if it overlapped completely with the previous one).

## Investigating signals of selection in modern humans

For the analysis of positive selection in modern humans, we used bi-allelic single nucleotide polymorphisms (SNPs) from 504 African individuals from the 1000 Genomes Project phase III (25), excluding African populations that may have European ancestry, *i.e.*, African Caribbeans in Barbados (ACB) and individuals with African Ancestry in Southwest US (ASW). We also compiled a list of sites where Africans differ from the common ancestor with chimpanzee. These are positions that are not polymorphic among the 504 African individuals and where six high-coverage African genomes (Mbuti, San and Yoruban individuals from (6)) were identical but differed from four great ape reference genome assemblies (panTro4, panPan1.1, gorGor3 and ponAbe2). We then extracted the genotypes of the *Altai Neandertal* (*Denisova 5*) or Denisovan (*Denisova 3*) genomes at these positions, and only considered those that pass the minimal filters for each genome, respectively (1).

We tested whether the missense substitutions in the spindle genes overlap regions displaying signatures of ancient selective sweeps using the hidden Markov model described in (24). We executed that method independently for each chromosome and for both the *Altai Neandertal* and *Denisova 3* genomes, using genetic distances computed from the African-American or deCODE recombination maps (26, 27). We then identified regions around the missense substitutions of spindle genes where the archaic genome fall outside the human variation (“external” regions) with a posterior probability above 0.2. We applied this cutoff on the sum of the posterior probabilities of both “external” states (*i.e.*, generated either from drift or from a selective sweep, (24)). We further intersected the regions called with the *Altai Neandertal* and *Denisova 3* genomes and measured the genetic length of the overlap to determine whether there is evidence for selection.

## Reconstructing the chronology of the missense substitutions

To get an upper age estimate of the missense substitutions, we estimated the TMRCA between Africans and archaic humans in the regions around the missense substitutions where archaic humans fall outside the human variation. We computed the number of pairwise differences between each African chromosome and each high-coverage archaic human genome (1, 5, 6, 28, 29) sampling a random allele at heterozygous positions in the archaic genomes. The age of the common ancestor of two individuals conditional on the number  $D$  of pairwise differences follows a gamma distribution with shape parameter  $\alpha=D+1$  and rate parameter  $\beta=\theta N+1$  (55), where  $\theta$  is the population scaled mutation rate (*i.e.*,  $4N_e\mu$  with  $\mu$  and  $N_e$  denoting the mutation rate and the effective population size, respectively) and  $N$  is the number of bases with genotype calls in both individuals. This model assumes that the prior probability of coalescence follows an exponential distribution, with rate equal to 1 when time is scaled in units of the diploid effective population size ( $2N_e$ ), and that both individuals are sampled in the present. To account for a branch shortening  $S$  associated with the age of ancient individuals, we truncated the posterior distribution so that the age of the common ancestor cannot be more recent than  $S$  and we added  $\frac{S}{2}$  as a correction for this branch shortening. The expected TMRCA is then:

$$E(T | T \geq S) = \frac{1}{\beta} \left( \alpha + \frac{\left(\frac{\beta S}{2}\right)^\alpha e^{-\frac{\beta S}{2}}}{\Gamma\left(\alpha, \frac{\beta S}{2}\right)} \right) + \frac{S}{2} \quad (\text{Equation 1, Appendix 6}), \text{ with } \Gamma\left(\alpha, \frac{\beta S}{2}\right) \text{ denoting the upper}$$

incomplete gamma function with lower limit  $\frac{\beta S}{2}$ , which we computed with the `gammainc()` function from the R library `expint` v0.1-6. We assumed  $\mu$  is  $1.45 \times 10^{-8}$  mutations per base pair per generation (generation time of 29 years, (56)) and a branch shortening of 50ky for *Vindija 33.19* (1), 70ky for *Denisova 3* (1), 80ky for *Chagyrkaya 8* (29), and 120ky for the *Altai Neandertal (Denisova 5)* (1). We computed the 95% confidence intervals with the `qtgamma()` function of the R library `TruncatedDistributions` v1.0. Note that we also estimated the TMRCA in these genomic regions between Africans and a Papuan (HGDP00548), a Sindhi (HGDP00163) or a Biaka (HGDP00461) because they carry the ancestral versions of the missense substitutions in *KATNA1*, *KIF18A* and *SPAG5*, respectively, and constrain further the upper age estimates of the missense substitutions.

To get a lower age estimate of the missense substitutions, we computed the number of pairwise differences among African chromosomes carrying the derived versions of the missense substitutions from the HGDP dataset (all African individuals, except HGDP00461 for *SPAG5*). For each region, we identified the two chromosomes that are the most distantly related (with the highest number of differences) and classified all the other African chromosomes into two groups depending on which of the two former chromosomes they are the most closely related in this genomic region. A chromosome was assigned to a group if it shared at least two derived alleles with the African chromosome used for defining this group but no more than one shared derived allele with the African chromosome defining the other group. This removes potential recombinants that could bias downward the estimate of the TMRCA. For each individual, we then counted the number of pairwise differences with individuals from the other group, restricting this analysis to positions that are genotyped in the high-coverage archaic genomes. Finally, we converted these counts into estimates of the TMRCA using the same equation as above, albeit with no branch shortening (*i.e.*,  $E(T) = \frac{\alpha}{\beta}$ ).

## Testing the hypothesis of gene flow in the *KNL1* gene

To test whether the *KNL1* haplotype identified in Neandertals could be shared with modern humans because of incomplete lineage sorting, we computed the probability that a haplotype shared since the common ancestral population is as long as, or longer than, the haplotype identified in *KNL1*. This approach was previously applied to the *EPAS1* haplotype shared between Tibetans and Denisovans (36), but we briefly describe it here for completeness. The expected length  $L$  for a shared sequence is  $1/(r \times T)$ , denoting  $r$  and  $T$  as the recombination rate and the length of the Neandertal and modern human branches since divergence (in generations of 29 years), respectively. As the ancestors of Neandertals and modern humans split from each other around 550kya (1), the most recent common ancestor shared between modern humans and Neandertals cannot be younger than this split time, in the absence of gene flow. As the modern-like *KNL1* haplotype is present in Neandertals that lived about 40kya (30), we set  $T = 550,000 - 40,000 = 510,000$  years, which corresponds to 17,586 generations of 29 years. Note that we do not include here the length of the modern human branch to be conservative, as we do not know when the substitutions that define the haplotype reached fixation in modern humans. Relying on local recombination rate estimates from the African-American map (0.148cM/Mb, (26)) and the deCODE map (0.191cM/Mb, (27)), the expected length  $L$  of a haplotype shared through ILS (*i.e.*,  $1/(r \times T)$ ) is 29.8kb or 38.5kb depending on the recombination map used. Assuming that the length of a shared sequence follows a Gamma distribution with shape parameter 2 and rate parameter  $1/L$ , the probability that such a sequence is as long as or longer than 276kb is then  $1 - \text{GammaCDF}(276000, \text{shape} = 2, \text{rate} = 1/L)$ . This probability is  $\leq 0.006$  and hence a model without gene flow is rejected.

## Estimating the time of gene flow

To estimate the time of gene flow, we used an analogous model to that described above ((55), adapted here to account for the branch shortening associated with the age of ancient individuals). In this model, the age of the common ancestor of two individuals conditional on the length  $L$  of their shared haplotype follows a gamma distribution with shape parameter  $\alpha=3$  and rate parameter  $\beta=2\rho L+1$ , where  $\rho$  is the population scaled recombination rate (*i.e.*,  $4N_e r$  with  $r$  and  $N_e$  denoting the recombination rate and the effective population size, respectively). To account for a branch shortening  $S$ , we applied again Equation 1, assuming that  $S$  is 40,000 years. In the case of the 276kb haplotype in Neandertals that carries alleles that are fixed in present-day humans, we made the conservative assumption that recombinations could not be observed in modern humans (*i.e.*, the alleles were already fixed at the time of introgression) and, therefore, multiplied the age estimates by 2. We did not apply this correction for the estimate based on the 102kb haplotype shared between some present-day humans and Neandertals, as the length of this haplotype depends on the number of recombination events in both Neandertals and modern humans. Using the average recombination rate over the African-American and deCODE recombination maps (0.169 cM/Mb), the expected age is 131ky based on the 276kb haplotype in Neandertals and 138ky based on the 102kb shared haplotype. We computed the 95% confidence intervals (83-265ky for the 276kb haplotype and 49-304ky for the 102kb haplotype) with the `qtgamma()` function of the R library `TruncatedDistributions` v1.0.

As another estimate for the time of gene flow, we look at the genetic distance between modern humans and Neandertals. For this purpose, we estimated the TMRCA between *Chagyrskaya 8* (the only high coverage genome with this haplotype) and each present-day human genome from the HGDP

dataset in the region chr15:40,885,107-40,963,160. We randomly sampled one allele at heterozygous positions where the phase is unknown in these genomes (*i.e.*, every heterozygous position of *Chagyrskaya 8*). The TMRCA conditional on the number  $D$  of pairwise differences follows a gamma distribution with parameters  $\alpha=D+1$  and  $\beta=\theta N+1$  (see dating analysis of the missense substitutions). However, conditional on both the observed divergence and the length of the shared haplotype, the TMRCA follows a gamma distribution with shape parameter  $\alpha=D+3$  and rate parameter  $\beta=\theta N+2pL+1$ . In both cases, we accounted for the branch shortening of *Chagyrskaya 8* as described above (Equation 1,  $S=80,000$  years) and assumed  $1.45 \times 10^{-8}$  mutations per base pair per generation (generation of 29 years). The 95% confidence intervals were computed as described above.

As these analyses might be sensitive to the value of  $\mu$ , we tested whether  $\mu$  in the region of *KNL1* may differ from the genome-wide average by computing local estimates of the mutation rate in family trios from Iceland (57). We counted the number of *de novo* mutations in the genomes of the probands and divided this number by twice the length of the region and the number of trios in this dataset (1,548) to get a mutation rate estimate. This estimate was similar to the genome-wide average (Appendix 7).

## Testing whether the modern-like *KNL1* haplotype was under selection in Neandertals

To test whether the increase of the modern-like *KNL1* haplotype frequency in Neandertals is unexpected, we quantified how often positions in the genome exhibit similar frequency changes. We identified positions where early Neandertals (the *Hohlenstein-Stadel*, *Scladina I-4A* and *Denisova 5* Neandertals) carry the ape-like allele, whereas at least one individual carries the derived allele both among those that lived 60-80kya (*Chagyrskaya 8*, *Mezmaiskaya 1*, *Denisova 11*) and among those that lived 40-50kya (*Vindija 33.19*, *Goyet Q56-1*, *Mezmaiskaya 2*, *Les Cottés Z4-1514* and *Spy 94a*). As noted above, only one random allele was considered for the low coverage genomes, in contrast to two for the high-coverage genomes. Positions with more than one missing allele among early Neandertals (out of 4 alleles) or 3 missing alleles (out of 6) among the late Neandertals were filtered out. We further removed transitions and positions less than 50kb away from the previously ascertained position. We then computed the proportion of those positions where at least three of the later individuals (*i.e.*, those that lived 40-50kya) carried the derived allele.

## Appendix 1

**Table 1: Location and predicted effects of the studied amino acid changes in spindle proteins, as reported in dbNSFP version 4.2 (48).** The predictions are for the ancestral variants. We put “damaging” in between quotation marks as the ancestral versions of ATRX and KATNA1 are unlikely to be damaging (as the ancestral amino acid residues are found in the proteins of many species), but that prediction rather supports a function for these amino acid changes.

protein	position	amino acid change	protein domain (Uniprot)	effect prediction for the ancestral variant (MutPred)	potentially “damaging” ancestral variant according to:
ATRX	475	D -> H	-	-	FATHMM; M-CAP
KATNA1	343	A -> T	-	-	FATHMM; PrimateAI
KIF18A	67	R -> K	kinesin motor (11-355)	Loss of methylation (p=0.0087)	-
KNL1	159	H -> R	interaction domain with BUB1 and BUB1B (1-728)	-	-
	1086	G -> S	2 X 104 AA approximate repeats (855-1201)	Loss of phosphorylation (p=0.0382)	-
NEK6	291	D -> H	protein kinase domain (45-310)	-	-
RSPH1	213	K -> Q	-	-	-
SPAG5	43	P -> S	-	Loss of phosphorylation (p=0.0244) Gain of catalytic residue (p=0.0179)	-
	162	E -> G	-	-	-
	410	D -> H	-	-	-
STARD9	3925	A -> T	-	-	-

561 **Table 2: Deleteriousness and conservation scores at the studied positions with missense changes in spindle genes, as reported in dbNSFP version 4.2 (48).** A high CADD score  
562 indicates that the ancestral variant is likely to be deleterious (58-60) and a high conservation score means that the nucleotide position is highly conserved across species (100  
563 vertebrates for phyloP and phastCons (61), and 34 mammals for GERP++ RS (62)). In contrast to the other scores that correspond to a single position, phastCons is a measure of  
564 the conservation in the region around the position. In dbNSFP, the scores range from -6.458163 to 18.301497 for CADD, from -12.3 to 6.17 for GERP++ RS, from -20.0 to 10.003  
565 for phyloP and from 0 to 1 for phastCons.

gene	position (hg19)	corresponding amino acid change	Deleteriousness	Conservation		
			CADD score (hg19)	GERP++ RS score	phyloP 100way vertebrate score	phastCons 100way vertebrate score
<i>KATNA1</i>	6-149,918,766	A343T	2.051033	5.48	4.834	1.000
<i>SPAG5</i>	17-26,925,570	P43S	-0.425670	-3.57	-1.404	0.000
	17-26,919,777	E162G	0.296317	3.66	0.280	0.001
	17-26,919,034	D410H	1.062743	5.4	2.032	1.000
<i>KNL1</i>	15-40,912,860	H159R	0.475454	3.7	-0.016	0.001
	15-40,915,640	G1086S	0.801787	4.12	1.026	0.054
<i>KIF18A</i>	11-28,119,295	R67K	1.134589	2.62	0.525	0.845
<i>NEK6</i>	9-127,113,155	D291H	-0.293112	-1.56	3.284	1.000
<i>ATRX</i>	X-76,939,325	D475H	-0.141606	3.64	0.791	0.840
<i>RSPH1</i>	21-43,897,491	K213Q	-0.061053	1.81	0.672	0.079
<i>STARD9</i>	15-42,985,549	A3925T	-0.351117	-2.51	0.047	0.000

566

## Appendix 2

**Table 1: Age estimates of the missense substitutions in spindle genes.** The ages were estimated in the regions where the *Altai Neandertal* and *Denisova 3* genomes fall outside the human variation (intersection of the regions identified with the African-American and deCODE recombination maps). The lower age corresponds to the mean age of the ancestor of multiple present-day African chromosome pairs. The upper age corresponds to the mean age of the common ancestor shared between each present-day African chromosome and either the archaic genome with the least number of differences (excluding *Chagyrskaya 8* for *KNL1*) or a present-day human with an ancestral version of the missense variant(s).

Gene	chromosome	Region (hg19)	Lower age (kya)	Upper age (kya)
<i>ATRX</i>	X	76,703,773-77,246,471	NA	NA
<i>KATNA1</i>	6	149,840,973-149,930,425	863	1,329
<i>KIF18A</i>	11	28,018,167-28,304,293	843	1,006
<i>KNL1</i>	15	40,898,141-40,948,306	1,027	1,690
<i>NEK6</i>	9	127,109,510-127,113,614	NA	NA
<i>RSPH1</i>	21	43,897,417-43,897,549	NA	NA
<i>SPAG5</i>	17	26,875,942-27,045,524	677	796
<i>STARD9</i>	15	42,941,540-42,989,160	947	1,401

# Appendix 3

**Table 1: Coverage depth in archaic human genomes at positions with modern human-specific missense substitutions in spindle genes.** The numbers of DNA fragments carrying a particular base are reported in parentheses after the corresponding base. Bases in uppercase were sequenced in the forward orientation, whereas those in lowercase were sequenced in the reverse orientation. Bases that are modern human-like (derived) are highlighted in red and may represent present-day human DNA contamination or an allele shared with modern humans. In contrast, the bases that are compatible with a damage-induced substitution (from the ancestral allele) are highlighted in blue (i.e., C-to-T and G-to-A in the forward and reverse orientation, respectively).

Gene	<i>ATRX</i>	<i>KATNA1</i>	<i>KIF18A</i>	<i>KNL1</i>		<i>NEK6</i>	<i>RSPH1</i>	<i>SPAG5</i>			<i>STARD9</i>
Chr-Position	X-76,939,325	6-149,918,766	11-28,119,295	15-40,912,860	15-40,915,640	9-127,113,155	21-43,897,491	17-26,919,034	17-26,919,777	17-26,925,578	15-42,985,549
Ancestral	C	C	C	A	G	G	T	C	T	G	G
Derived	G	T	T	G	A	C	G	G	C	A	A
<i>Altai Neandertal</i>	C (21) c (31)	C (22) c (15) <b>T (2)</b>	C (19) c (41) <b>T (1)</b>	A (24) a (28)	G (33) g (16)	G (23) g (23)	T (18) t (26)	C (17) c (17)	T (18) t (20) a (1)	G (28) g (17)	G (13) g (16) <b>A (1)</b>
<i>Chagyrskaya 8</i>	C (10) c (9) <b>T (1)</b> t (1)	C (13) c (5)	C (15) c (19) <b>T (3)</b>	A (1) <b>G (14)</b> g (11)	<b>A (19)</b> a (8)	G (5) g (7) a (1)	T (17) t (13)	C (11) c (7) <b>T (1)</b>	T (16) t (11)	G (7) g (6) a (1)	G (8) g (13)
<i>Denisova 3</i>	C (19) c (17)	C (15) c (13) <b>T (2)</b>	C (17) c (24)	A (20) a (30)	G (25) g (17) a (1)	G (15) g (14)	T (19) t (27)	C (20) c (14)	T (16) t (11)	G (12) g (6)	G (17) g (12)
<i>Denisova 11</i>	C (1) c (2)	NA	C (1)	a (2)	g (1)	G (1) g (1)	T (4) t (2)	c (2) t (1)	NA	NA	G (1)
<i>Goyet Q56-1</i>	C (1)	NA	C (1) c (5)	<b>G (1)</b>	<b>A (1)</b> a (2)	G (1) g (2)	NA	C (1)	T (1)	G (4) g (2)	G (2)
<i>Hohlenstein-Stadel</i>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
<i>Les Cottés Z4-1514</i>	C (1) c (1)	c (2) <b>T (1)</b>	C (5) c (4)	A (2) a (4)	G (3) g (2)	NA	T (2) t (1)	NA	T (1)	G (1)	G (1)
<i>Mezmaiskaya 1</i>	c (3)	NA	C (1) c (1)	NA	G (2)	NA	t (2)	C (2) c (1) g (2)	NA	NA	NA
<i>Mezmaiskaya 2</i>	C (1)	C (1) c (1) <b>T (1)</b>	C (1)	<b>g (1)</b>	<b>A (1)</b>	G (2) g (1)	NA	C (1)	T (1)	G (2)	g (3)
<i>Scladina I-4A</i>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
<i>Spy 94A</i>	NA	c (1)	C (1)	<b>g (1)</b>	<b>A (1)</b> a (3)	NA	T (1)	C (1)	T (1)	g (1)	NA
<i>Vindija 33.19</i>	C (16) c (18) <b>T (1)</b>	C (8) c (11)	C (17) c (20) <b>T (1)</b>	A (13) a (19)	G (20) g (16) a (2)	G (8) g (8)	T (22) t (15)	C (12) c (14)	T (15) t (7)	G (15) g (13) a (1)	G (14) g (14)

**Table 2: Coverage depth of the *Mezmaiskaya 1* genome at positions with modern human-specific substitutions in *SPAG5*.** Only positions covered by at least one DNA sequence are reported. Bases in uppercase were sequenced in the forward orientation, whereas those in lowercase were sequenced in the reverse orientation. The numbers of DNA fragments carrying a particular base are reported in parentheses after the corresponding base.

Neandertal	Chr-position	Ancestral	Derived	Allele counts
<i>Mezmaiskaya 1</i>	17-26,864,608	A	G	A (1)
	17-26,891,162	T	G	T (1)
	17-26,892,376	A	T	A (2)
	17-26,913,024	A	G	a (a)
	17-26,919,034	C	G	C (2) c (1) g (2)
	17-26,948,236	G	A	g (1)
	17-26,967,723	A	G	A (3)
	17-27,005,275	G	A	G (1)
	17-27,010,483	G	A	g (1)

## Appendix 4

**Table 1: Positions defining the closely related haplotype between some modern humans and Neandertals.** At these positions, the *Chagyrskaya 8* genome differs from other high-quality archaic genomes without the modern human-like haplotype but some African genomes from the HGDP dataset carry the same allele as *Chagyrskaya 8*. Note that the modern human-like haplotype in Neandertals is longer and defined by alleles that are shared with all modern humans (Figure 3).

Chromosome	Position (hg19)	Reference	Alternative ( <i>Chagyrskaya 8</i> -like)	<i>Chagyrskaya 8</i> -like allele frequency in genomes from the HGDP dataset	
				Africans	Non-Africans
15	40,885,107	A	G	0.32	0.41
	40,886,017	C	T	0.32	0.41
	40,886,020	C	G	0.32	0.40
	40,892,601	G	A	0.35	0.40
	40,893,573	C	A	0.38	0.40
	40,905,450	C	T	0.37	0.41
	40,908,904	A	G	0.39	0.41
	40,910,707	T	C	0.38	0.41
	40,915,045	T	G	0.38	0.41
	40,915,894	T	C	0.39	0.41
	40,925,214	T	A	0.39	0.41
	40,926,654	C	G	0.35	0.40
	40,929,814	T	C	0.37	0.41
	40,937,647	A	G	0.38	0.41
	40,959,413	G	A	0.36	0.10
	40,959,624	G	A	0.36	0.40
	40,960,432	G	A	0.36	0.40
	40,963,160	A	G	0.37	0.41
	40,987,528	G	C	0.38	0.11

## Appendix 5

**Table 1: Allele counts at positions with nearly fixed missense variants in the spindle genes of modern humans from the *gnomad* database (v2.1.1)(38).** Columns 7-8 and 9-10 correspond to the allele counts among the 125,748 exome sequences and the 15,708 whole-genome sequences, respectively. Anc= Ancestral

Gene	Chr-Position	Anc	(nearly) fixed	Alleles	VEP An-not.	# Anc	Total	# Anc	Total
<i>ATRX</i>	X-76,939,325	C	G	G-C	missense	66	182,745	11	22,042
<i>KATNA1</i>	6-149,918,766	C	T	T-C	missense	259	251,190	131	31,400
<i>KIF18A</i>	11-28,119,295	C	T	T-C	missense	26	249,508	1	31,396
<i>KNL1</i>	15-40,912,860	A	G	G-A	missense	NA	NA	1	31,368
				G-T	missense	1	227,420	NA	NA
	15-40,915,640	G	A	A-G	missense	NA	NA	NA	NA
<i>NEK6</i>	9-127,113,155	G	C	C-G	missense	164	250,140	26	31,404
<i>RSPH1</i>	21-43,897,491	T	G	G-T	missense	236	251,414	30	31,386
				G-A	stop gained	10	251,414	1	31,386
<i>SPAG5</i>	17-26,919,034	C	G	G-C	missense	NA	NA	NA	NA
	17-26,919,777	T	C	C-A	missense	3	251,430	NA	NA
	17-26,925,578	G	A	A-G	missense	NA	NA	NA	NA
				A-T	missense	1	251,066	NA	NA
<i>STARD9</i>	15-42,985,549	G	A	A-G	missense	5	139,342	3	31,284
				A-C	missense	5	139,342	NA	NA

## Appendix 6: Derivation of Equation 1

To compute the Time to the Most Recent Common Ancestor (TMRCA) for pairs of modern and archaic humans, we used a published model (55) that we adapted to account for the branch shortening associated with the age of the archaic individual. We simply truncated the posterior distribution of the TMRCA obtained with this model so that the TMRCA cannot be more recent than the age of the archaic individual, and added a correction to account for missing mutations on the archaic branch. Here, we describe how we derived Equation 1 to compute the expected TMRCA with these modifications from the original model.

The expectation of the truncated distribution is:

$$\begin{aligned} E(X | X \geq T) &= \frac{E(X \cap X \geq T)}{P(X \geq T)}, \text{ with } X \text{ denoting the time to coalescence and } T \text{ the truncated time,} \\ &= \frac{\int_T^\infty x \text{Gamma}(x | \alpha, \beta) dx}{\int_T^\infty \text{Gamma}(x | \alpha, \beta) dx}, \text{ with } \text{Gamma}(x | \alpha, \beta) \text{ denoting the probability density function of the gamma} \\ &\text{distribution (with parameters } \alpha \text{ and } \beta \text{ defined as in (55)),} \\ &= \frac{\int_T^\infty x^\alpha e^{-\beta x} dx}{\int_T^\infty x^{\alpha-1} e^{-\beta x} dx} \\ &= \frac{1}{\beta} \frac{\int_T^\infty (\beta x)^\alpha e^{-\beta x} dx}{\int_T^\infty (\beta x)^{\alpha-1} e^{-\beta x} dx} \\ &= \frac{1}{\beta} \frac{\int_{\beta T}^\infty y^\alpha e^{-y} dy}{\int_{\beta T}^\infty y^{\alpha-1} e^{-y} dy} \\ &= \frac{1}{\beta} \frac{\Gamma(\alpha+1, \beta T)}{\Gamma(\alpha, \beta T)}, \text{ with } \Gamma(\alpha, \beta T) \text{ denoting the upper incomplete gamma function.} \end{aligned}$$

By integration by parts, one can show that  $\Gamma(\alpha + 1, \beta T) = \alpha \Gamma(\alpha, \beta T) + (\beta T)^\alpha e^{-\beta T}$  to obtain:

$$E(X | X \geq T) = \frac{1}{\beta} \left( \alpha + \frac{(\beta T)^\alpha e^{-\beta T}}{\Gamma(\alpha, \beta T)} \right)$$

As this model assumes that the two lineages diverge for the same amount of time (*i.e.*, for a given coalescent time  $T$  the total branch length is  $2T$ ), we therefore set  $T = \frac{S}{2}$ , with  $S$  denoting the age of the ancient specimen, and added  $\frac{S}{2}$  as a correction for the branch shortening (*i.e.*, adds  $S$  to the total branch length). This leads to Equation 1:

$$E(\text{TMRCA} | \text{TMRCA} \geq S) = \frac{1}{\beta} \left( \alpha + \frac{\left(\frac{\beta S}{2}\right)^\alpha e^{-\frac{\beta S}{2}}}{\Gamma(\alpha, \frac{\beta S}{2})} \right) + \frac{S}{2}$$

## Appendix 7: Local mutation rate in the region of *KNL1*

For estimating the age of common ancestors shared between modern humans and Neandertals in the region of *KNL1*, we assumed the genome-wide average of  $1.45 \times 10^{-8}$  mutations per base pair per generation (*i.e.*, the mutation rate used to estimate the population split times between the ancestors of modern humans and Neandertals (1, 6); estimated from (56)). As the age estimates may be sensitive to the mutation rate used, we also estimated the local mutation rate among family trios from Iceland (57). We counted the number of *de novo* mutations in the genomes of the probands and divided this number by twice the length of the region and the number of trios in this dataset (1,548) and arrive at a mutation rate estimate of  $0.94$  [95% Binomial CI:  $0.4-1.8$ ]  $\times 10^{-8}$  mutations per base pair per generation in the *KNL1* region where modern humans and *Chagyrskaya 8* share a haplotype. However, there were only 8 *de novo* mutations among the Icelandic genomes in this region. If we extend that region by 500kb on both sides to increase the accuracy of the estimate, the mutation rate is  $1.1$  [95% Binomial CI:  $0.9-1.6$ ]  $\times 10^{-8}$  mutations per base pair per generation (45 *de novo* mutations in this extended region), which is similar to the genome-wide average mutation rate of  $1.29 \times 10^{-8}$  estimated in the original study of the trios (57). Therefore, there is no evidence that the mutation rate is different from the genome-wide average in this particular region of the genome.

# Acknowledgments

We thank Felipe Mora-Bermúdez, Wieland Huttner, Divyaratan Popli and Hugo Zeberg for helpful discussions and Alba Bossoms Mesa and Leonardo N. Iasi for suggestions for the figures. This work was supported by the Max Planck Society, the European Research Council [694707] and the NOMIS Foundation.

# Supplement File 1

**Genomic regions where archaic humans fall outside the modern human variation, identified using the most recent deCODE recombination map (27).**

# Supplement File 2

**Genotypes of the 12 non-African individuals that inherited one copy of *KNL1* from archaic humans.** We show positions within 40kb downstream of the modern-like *KNL1* haplotype identified in *Chagyrskaya 8* to highlight 7 positions (red marks) where only those 12 individuals (out of 929 individuals across worldwide populations (28)) carry a derived allele seen in at least one Neandertal genome. The upper panel (“Archaics”) shows the alleles carried by high-coverage archaic human genomes without the modern-like *KNL1* haplotype. The middle panel (“*Chagyrskaya 8*-like”) shows the alleles carried by four Neandertal genomes with the modern-human like *KNL1* haplotype and the 12 present-day non-African genomes that inherited one copy of *KNL1* from archaic humans. The lower panel (“Other haplotypes”) shows the alleles carried by the other chromosomes (that did not inherit a copy of *KNL1* from archaic humans) of these 12 individuals. For the archaic human genomes, one allele was sampled randomly at heterozygous positions.

# References

1. K. Prüfer *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658 (2017).
2. Erik Trinkaus, Modern Human versus Neandertal Evolutionary Distinctiveness. *Current Anthropology* **47**, 597-620 (2006).
3. W. Roebroeks, M. Soressi, Neandertals revised. *Proc Natl Acad Sci U S A* **113**, 6372-6379 (2016).
4. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).
5. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).

- 684 6. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains.  
685 *Nature* **505**, 43-49 (2014).
- 686 7. G. J. Mack, D. A. Compton, Analysis of mitotic microtubule-associated proteins using mass  
687 spectrometry identifies astrin, a spindle-associated protein. *Proc Natl Acad Sci U S A* **98**, 14434-  
688 14439 (2001).
- 689 8. M. S. Chang *et al.*, Cloning and characterization of hMAP126, a new member of mitotic spindle-  
690 associated proteins. *Biochem Biophys Res Commun* **287**, 116-121 (2001).
- 691 9. S. L. Prosser, L. Pelletier, Mitotic spindle assembly in animal cells: a fine balancing act. *Nat Rev*  
692 *Mol Cell Biol* **18**, 187-201 (2017).
- 693 10. I. M. Cheeseman, T. Hori, T. Fukagawa, A. Desai, KNL1 and the CENP-H/I/K complex coordinately  
694 direct kinetochore assembly in vertebrates. *Mol Biol Cell* **19**, 587-594 (2008).
- 695 11. M. I. Mayr *et al.*, The human kinesin Kif18A is a motile microtubule depolymerase essential for  
696 chromosome congression. *Curr Biol* **17**, 488-498 (2007).
- 697 12. S. Paabo, The human condition-a molecular approach. *Cell* **157**, 216-226 (2014).
- 698 13. M. Kuhlwilm, C. Boeckx, A catalog of single nucleotide changes distinguishing modern humans  
699 from archaic hominins. *Sci Rep* **9**, 8463 (2019).
- 700 14. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology  
701 Consortium. *Nat Genet* **25**, 25-29 (2000).
- 702 15. C. Gene Ontology, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**,  
703 D325-D334 (2021).
- 704 16. K. Ritchie, L. A. Watson, B. Davidson, Y. Jiang, N. G. Berube, ATRX is required for maintenance of  
705 the neuroprogenitor cell pool in the embryonic mouse brain. *Biol Open* **3**, 1158-1163 (2014).
- 706 17. M. Ramamoorthy, S. Smith, Loss of ATRX Suppresses Resolution of Telomere Cohesion to  
707 Control Recombination in ALT Cancer Cells. *Cancer Cell* **28**, 357-369 (2015).
- 708 18. A. Roll-Mecak, F. J. McNally, Microtubule-severing enzymes. *Curr Opin Cell Biol* **22**, 96-103  
709 (2010).
- 710 19. M. J. Yin, L. Shao, D. Voehringer, T. Smeal, B. Jallal, The serine/threonine kinase Nek6 is required  
711 for cell cycle progression through mitosis. *J Biol Chem* **278**, 52454-52460 (2003).
- 712 20. G. Vaz Meirelles *et al.*, Characterization of hNek6 interactome reveals an important role for its  
713 short N-terminal domain and colocalization with proteins at the centrosome. *J Proteome Res* **9**,  
714 6298-6316 (2010).
- 715 21. J. Tsuchida *et al.*, Molecular cloning and characterization of meichroacidin (male meiotic  
716 metaphase chromosome-associated acidic protein). *Dev Biol* **197**, 67-76 (1998).
- 717 22. J. Z. Torres *et al.*, The STARD9/Kif16a kinesin associates with mitotic microtubules and regulates  
718 spindle pole assembly. *Cell* **147**, 1309-1323 (2011).
- 719 23. R. R. Hudson, M. Kreitman, M. Aguade, A test of neutral molecular evolution based on  
720 nucleotide data. *Genetics* **116**, 153-159 (1987).
- 721 24. S. Peyregne, M. J. Boyle, M. Dannemann, K. Prüfer, Detecting ancient positive selection in  
722 humans using extended lineage sorting. *Genome Res* **27**, 1563-1572 (2017).
- 723 25. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74  
724 (2015).
- 725 26. A. G. Hinch *et al.*, The landscape of recombination in African Americans. *Nature* **476**, 170-175  
726 (2011).
- 727 27. B. V. Halldorsson *et al.*, Characterizing mutagenic effects of recombination through a sequence-  
728 level genetic map. *Science* **363** (2019).
- 729 28. A. Bergstrom *et al.*, Insights into human genetic variation and population history from 929  
730 diverse genomes. *Science* **367** (2020).

- 731 29. F. Mafessoni *et al.*, A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of*  
732 *the National Academy of Sciences* **117**, 15132 (2020).
- 733 30. M. Hajdinjak *et al.*, Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652-656  
734 (2018).
- 735 31. S. Peyrégne *et al.*, Nuclear DNA from two early Neanderthals reveals 80,000 years of genetic  
736 continuity in Europe. *Sci Adv* **5**, eaaw5873 (2019).
- 737 32. V. Slon *et al.*, The genome of the offspring of a Neanderthal mother and a Denisovan father.  
738 *Nature* **561**, 113-116 (2018).
- 739 33. A. W. Briggs *et al.*, Patterns of damage in genomic DNA sequences from a Neanderthal. *Proc Natl*  
740 *Acad Sci U S A* **104**, 14616-14621 (2007).
- 741 34. H. Zeberg, S. Paabo, The major genetic risk factor for severe COVID-19 is inherited from  
742 Neanderthals. *Nature* **587**, 610-612 (2020).
- 743 35. M. Dannemann, A. M. Andres, J. Kelso, Introgression of Neanderthal- and Denisovan-like  
744 Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am J Hum Genet* **98**,  
745 22-33 (2016).
- 746 36. E. Huerta-Sanchez *et al.*, Altitude adaptation in Tibetans caused by introgression of Denisovan-  
747 like DNA. *Nature* **512**, 194-197 (2014).
- 748 37. L. N. M. Iasi, H. Ringbauer, B. M. Peter, An Extended Admixture Pulse Model Reveals the  
749 Limitations to Human-Neanderthal Introgression Dating. *Mol Biol Evol* 10.1093/molbev/msab210  
750 (2021).
- 751 38. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456  
752 humans. *Nature* **581**, 434-443 (2020).
- 753 39. F. Mora-Bermudez *et al.*, Differences and similarities between human and chimpanzee neural  
754 progenitors during cerebral cortex development. *Elife* **5** (2016).
- 755 40. F. Mora-Bermudez, E. Taverna, W. B. Huttner, From stem and progenitor cells to neurons in the  
756 developing neocortex: key differences among hominids. *FEBS J* 10.1111/febs.15793 (2021).
- 757 41. M. Meyer *et al.*, Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos  
758 hominins. *Nature* **531**, 504-507 (2016).
- 759 42. C. Posth *et al.*, Deeply divergent archaic mitochondrial genome provides lower time boundary  
760 for African gene flow into Neanderthals. *Nat Commun* **8**, 16046 (2017).
- 761 43. M. Petr *et al.*, The evolutionary history of Neanderthal and Denisovan Y chromosomes. *Science*  
762 **369**, 1653-1656 (2020).
- 763 44. M. Kuhlwilm *et al.*, Ancient gene flow from early modern humans into Eastern Neanderthals.  
764 *Nature* **530**, 429-433 (2016).
- 765 45. M. J. Hubisz, A. L. Williams, A. Siepel, Mapping gene flow between ancient hominins through  
766 demography-aware inference of the ancestral recombination graph. *PLoS Genet* **16**, e1008895  
767 (2020).
- 768 46. K. Harvati *et al.*, Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia.  
769 *Nature* **571**, 500-504 (2019).
- 770 47. I. Herskovitz *et al.*, The earliest modern humans outside Africa. *Science* **359**, 456-459 (2018).
- 771 48. X. Liu, C. Li, C. Mou, Y. Dong, Y. Tu, dbNSFP v4: a comprehensive database of transcript-specific  
772 functional predictions and annotations for human nonsynonymous and splice-site SNVs.  
773 *Genome Med* **12**, 103 (2020).
- 774 49. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
- 775 50. S. Chimpanzee, C. Analysis, Initial sequence of the chimpanzee genome and comparison with the  
776 human genome. *Nature* **437**, 69-87 (2005).
- 777 51. K. Prufer *et al.*, The bonobo genome compared with the chimpanzee and human genomes.  
778 *Nature* **486**, 527-531 (2012).

52. A. Scally *et al.*, Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175 (2012).
53. D. P. Locke *et al.*, Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533 (2011).
54. R. S. Harris, *Improved pairwise alignment of genomic DNA* (The Pennsylvania State University, 2007).
55. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol* **18**, e3000586 (2020).
56. A. Scally, R. Durbin, Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13**, 745-753 (2012).
57. H. Jonsson *et al.*, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).
58. M. Kircher *et al.*, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).
59. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
60. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010).
61. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).
62. E. V. Davydov *et al.*, Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).

