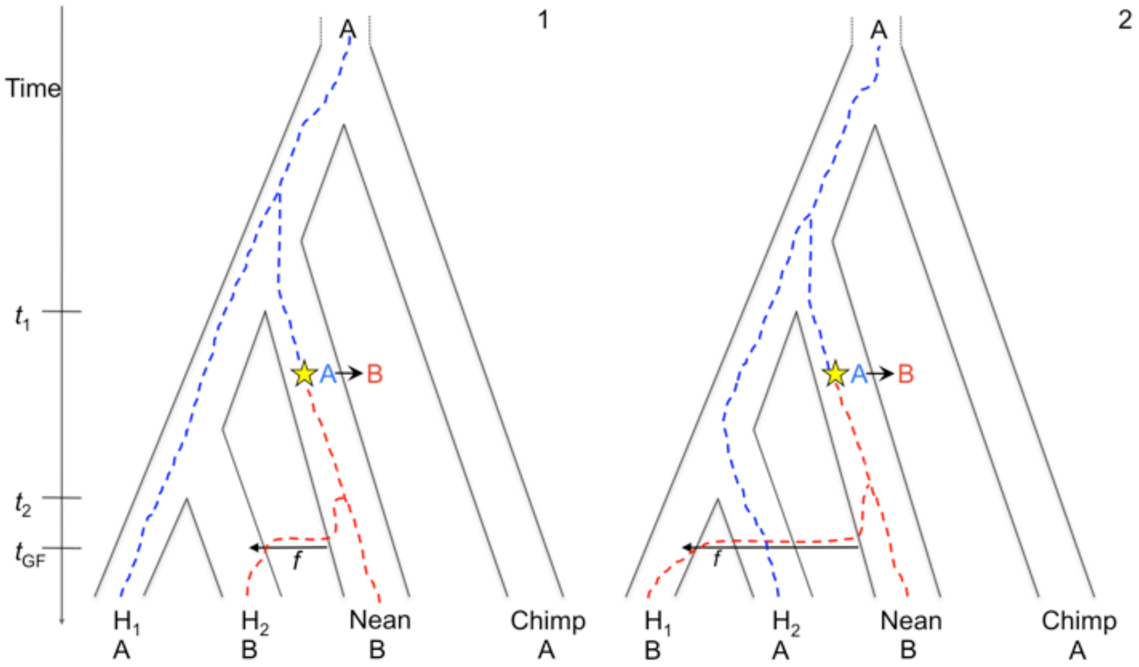
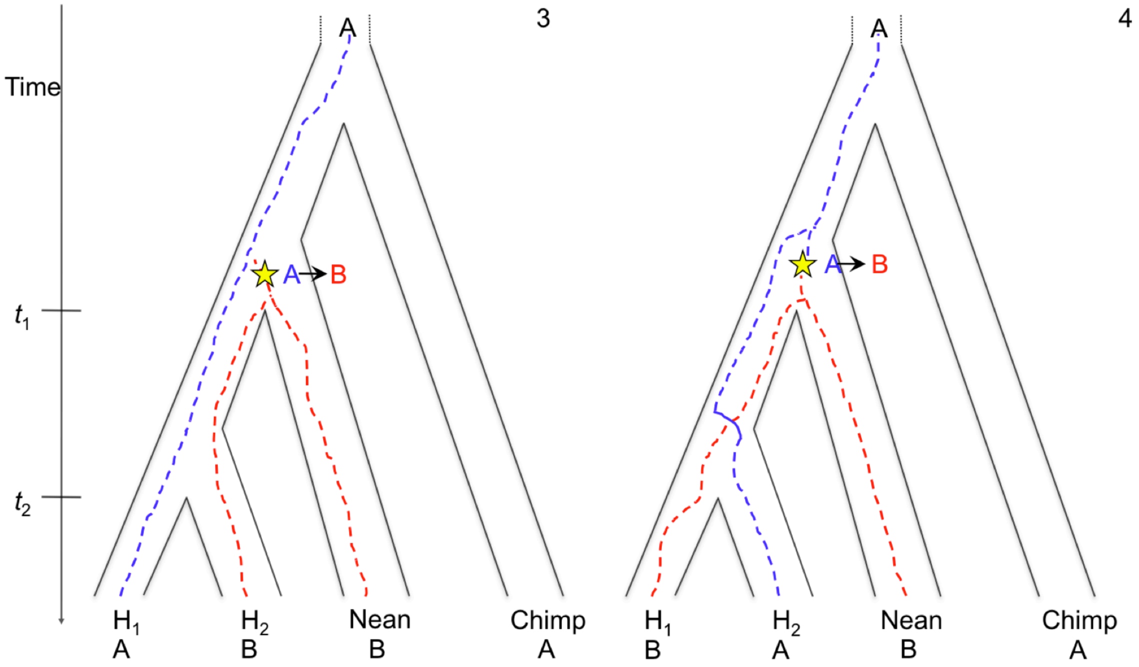


$$D(H_1,H_2,Nean,Chimp)=\frac{\sum_{i=1}^{L} I_{ABBA}(i)-I_{BABA}(i)}{\sum_{i=1}^{L} I_{ABBA}(i)+I_{BABA}(i)}$$



BOX S1 statistic. The D statistic[1, 2] measures the rate of differential allele sharing between an archaic human population, here the Neanderthal (Nean.), to 2 modern human populations, H_1 and H_2. The comparison is made in the simplest case between 4 haplotypes, one from each of the groups, and our formula refers to that case. The ancestral allele, A, is defined by the chimpanzee sequence (Chimp) and the derived allele is labeled B. We only consider sites for which the Neanderthal has the derived allele. We then count how many sites have a configuration (H1,H2,Nean,Chimp) = (Ancestral, Derived, Derived, Ancestral) = "ABBA" versus (H1,H2,Nean,Chimp) = (Derived, Ancestral, Derived, Ancestral) = "BABA". These two numbers are subtracted and appropriately normalized so the statistic lies between -1 and 1. A significant excess of ABBA sites over BABA sites, $D > 0$, would be indicative of more relatedness of H_2 to the Neanderthal, while an excess of BABA over ABBA sites, $D < 0$, would indicate more relatedness of H_1 to the Neanderthal. We consider the genealogies in which we would observe ABBA or BABA sites. In the absence of Neanderthal admixture due to lineage sorting, in tree 3 or tree 4, we would need a derived polymorphism (A-to-B) to arise in the population ancestral to H_1, H_2 and Nean. This derived polymorphic allele will be observed equally likely in the H_2 haplotype leading to an ABBA site (tree 3) or the H_1 haplotype producing a BABA site (tree 4), and so we will see $D \approx 0$ on average. In the presence of admixture (see Tree 1 or 2), in addition to the class of sites produced by trees 3 and 4, A-to-B mutations that have occurred on the Neanderthal branch will lead to an excess of ABBA or BABA sites, the former when admixture occurs in H_2 ($D>0$), and the latter when admixture is in H1 ($D<0$). While this example assumes that the chimpanzee allele matches the ancestral allele (i.e. the allele from the ancestor of H_1, H_2 and Neanderthal), if the two alleles differ then we will not observe an ABBA or BABA configuration at all (i.e. if the mutation occurred in the chimp branch or in the ancestral population to H_1, H_2 and Neanderthal).

BOX S2 – Using haplotype length to distinguish introgression from ancestral shared polymorphism

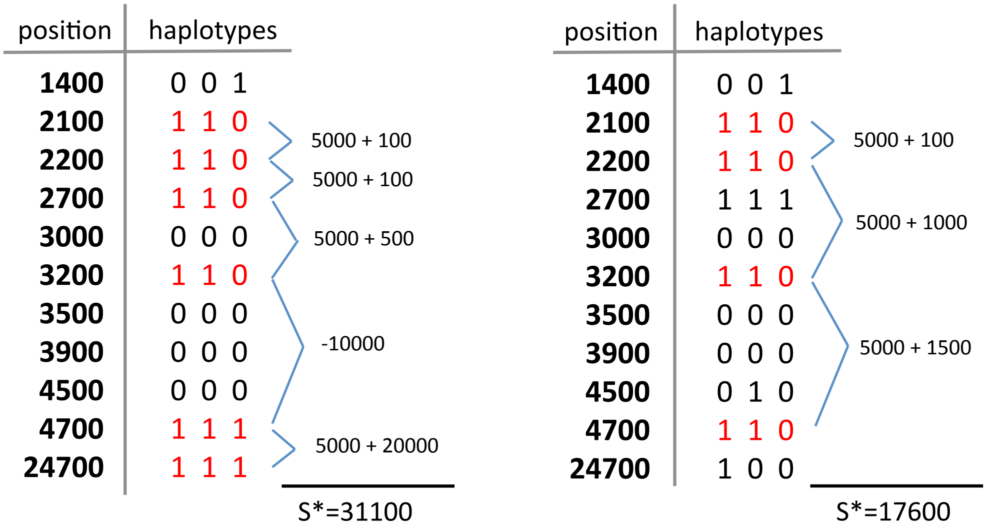
A high level of nucleotide similarity between sequences found in high linkage disequilibrium in a specific population and an outgroup population or species need not be caused by introgression from the outgroup. An alternative explanation is ancestral shared polymorphism, perhaps caused by deep population structure. This model posits that a haplotype existed before the divergence from the outgroup, and only survived in a particular population and the outgroup, but not in other populations (Figure 1, Box 1 panel 2 and 3). A way to distinguish ancestral shared haplotypes from introgressed haplotypes is by measuring their length[3-5].

Recombination breaks apart haplotypes over time, and so long shared haplotypes between any two sets of lineages tend to be recent. In turn, short-shared haplotypes tend to be older and are therefore more likely to be caused by ancestral shared polymorphism (Figure 1). If the lengths of introgressed tracts are exponentially distributed, then the probability of observing a haplotype of length >= *k* that is shared between two populations approximately follows an exponential distribution with parameter k/L[6], where L is the expected length of a shared sequence (but see Liang and Nielsen 2014[7] for violations of this assumption).

P(block size >= k) = exp(-k/L)

The parameter L is equal to $([1-m]r[t-1])^{-1}$ $[\approx 1/(r\ t)]$, where t is the time separating the two populations (in generations), r is the recombination rate per bp per generation and m is the admixture fraction. The above formula is valid for randomly chosen introgression tracts. However, for randomly chosen sites in the genome, conditionally on the site containing introgressed DNA, the length of the tract will be a sum of two exponential random variables (each one representing the sequence length to the left and the right from the chosen site). This model does not include selection, and it does not depend on the frequency of the haplotype.

For example, let there be a haplotype of length 40,000 bp that is shared between two populations that split 8,000 generations in the past. This haplotype is in a region with recombination rate = 2.5e-8. We can calculate that the expected length of a shared sequence to be approximately L = 5,000 bp, and so the haplotype has a probability of 0.0003 of having persisted due to ancestral shared polymorphism.



position	haplotypes
1400	0 0 1
2100	1 1 0
2200	1 1 0
2700	1 1 0
3000	0 0 0
3200	1 1 0
3500	0 0 0
3900	0 0 0
4500	0 0 0
4700	1 1 1
24700	1 1 1

position	haplotypes
1400	0 0 1
2100	1 1 0
2200	1 1 0
2700	1 1 1
3000	0 0 0
3200	1 1 0
3500	0 0 0
3900	0 0 0
4500	0 1 0
4700	1 1 0
24700	1 0 0

$$S(i,j) = \begin{cases} -\text{Inf if } j \text{ is a singleton} \\ -\text{Inf if } |i-j| < 10 \text{ or } |i-j| > 50000 \\ 5000 + |i-j| \text{ if } d(i,j) = 0 \text{ (no mismatches)} \\ -10000 \text{ if } 1 \leq d(i,j) \leq 5 \text{ (up to 5 mismatches)} \end{cases}$$

d(i,j) = Number of haplotypes that do not match on SNP i and j

BOX S3. Two examples of S* calculations (using haplotype data for simplicity). S* is calculated by optimizing over all subsets of SNPs at a given locus the sum of scores *S(i,j)* where i and j are two successive (not necessarily adjacent) SNP positions in a given subset[8-11]. The form of *S(i,j)* is heuristic but it rewards fully linked pairs of sites (d(i,j)=0 where d(i,j) is the number of haplotypes that differ between position i and j) and the reward is increased in proportion to the distance between the positions. Therefore it tends to discover subsets of SNPs within a window that are in high linkage disequilibrium (LD), and the S* value is higher the higher the LD. Twelve positions (coordinates are in bp) and three haplotypes are shown in each example. Sites labeled 0 are ancestral (like chimpanzee) and sites labeled 1 are derived. The S* calculation proceeds via a dynamic programming algorithm, and as such guarantees that the optimal solution (in red) gives the maximal value for the sum of *S(i,j)* values, i.e. there is no other selection of SNPs that will give a higher sum of *S(i,j)* in each of these examples. In the example on the left, position 2700 has higher linkage to its neighboring SNPs than in the example on the right, and therefore this site is included in the haplotype. Additionally, position 4700 and 24700 are linked and their distance is rewarded by S(i,j), and outweighs the penalty incurred from position 3200 to 4700.

References

1. Green, R.E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710-22 (2010).
2. Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239-2252 (2011).
3. Mendez, F.L., Watkins, J.C. & Hammer, M.F. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am J Hum Genet* **91**, 265-74 (2012).
4. Ding, Q., Hu, Y., Xu, S., Wang, J. & Jin, L. Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. *Mol Biol Evol* **31**, 683-95 (2014).
5. Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-7 (2014).
6. Pool, J.E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711-9 (2009).
7. Liang, M. & Nielsen, R. The Lengths of Admixture Tracts. *Genetics* **197**, 953-967 (2014).
8. Plagnol, V. & Wall, J.D. Possible ancestral structure in human populations. *PLoS Genet* **2**, e105 (2006).
9. Wall, J.D., Lohmueller, K.E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol* **26**, 1823-7 (2009).
10. Wall, J.D. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154**, 1271-1279 (2000).
11. Vernot, B. & Akey, J.M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-21 (2014).