Genetics: Early Online, published on February 14, 2013 as 10.1534/genetics.112.148213

Higher levels of Neanderthal ancestry in East Asians than in Europeans

Jeffrey D. Wall[*§], Melinda A. Yang[†], Flora Jay[†], Sung K. Kim*[1], Eric Y. Durand[2†], Laurie S. Stevison[*], Christopher Gignoux[*], August Woerner[‡], Michael F. Hammer[‡] and Montgomery Slatkin[†]

[*] Institute for Human Genetics, University of California, San Francisco, CA 94143

[§] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143

[†] Department of Integrative Biology, University of California, Berkeley, CA 94720

[‡] Department of Arizona Research Laboratories, University of Arizona, Tucson, AZ 85721

[1] Current Address: Sequenom, Inc., San Diego, CA 92121

[2] Current Address: 23andMe, Mountain View, CA 94043

Running title: Neanderthal ancestry in Eurasians

Key words: Human evolution, Neanderthals, ancient admixture

Correspondence to:

Jeff Wall

513 Parnassus Avenue, S965

San Francisco, CA 94143

Phone: (415)-476-4063

Email: wallj@humgen.ucsf.edu

Abstract

Neanderthals were a group of archaic hominins that occupied most of Europe and parts of Western Asia from roughly 30 – 300 thousand years ago (Kya). They coexisted with modern humans during part of this time. Previous genetic analyses that compared a draft sequence of the Neanderthal genome with genomes of several modern humans concluded that Neanderthals made a small (1-4%) contribution to the gene pools of all non-African populations. This observation was consistent with a single episode of admixture from Neanderthals into the ancestors of all non-Africans when the two groups coexisted in the Middle East 50 – 80 Kya. We examined the relationship between Neanderthals and modern humans in greater detail by applying two complementary methods to the published draft Neanderthal genome and an expanded set of high-coverage modern human genome sequences. We find that, consistent with the recent finding of Meyer et al. (2012), Neanderthals contributed more DNA to modern East Asians than to modern Europeans. Furthermore we find that the Maasai of East Africa have a small but significant fraction of Neanderthal DNA. Because our analysis is of several genomic samples from each modern human population considered, we are able to document the extent of variation in Neanderthal ancestry within and among populations. Our results combined with those previously published show that a more complex model of admixture between Neanderthals and modern humans is necessary to account for the different levels of Neanderthal ancestry among human populations. In particular, at least some Neanderthal – modern human admixture must postdate the separation of the ancestors of modern European and modern East Asian populations.

INTRODUCTION

Neanderthals were a group of archaic hominins that occupied large parts of Europe and West Asia from roughly 30 – 300 thousand years ago (Kya) (HUBLIN 2009; STRINGER and HUBLIN 1999). Their disappearance in the fossil record often coincides with the first appearance of anatomically modern humans (AMH) in that region (FINLAYSON 2004). Where, when, and how often Neanderthals interbred with expanding AMH populations is still an open question. Morphological studies have generally concluded that Neanderthals made little or no contribution to present-day human populations (LAHR 1994; STRINGER and ANDREWS 1988), but others have suggested there was some admixture (DUARTE *et al.* 1999; TRINKAUS 2007). Initial comparisons of Neanderthal and modern human DNA found no evidence for a Neanderthal contribution to the modern human gene pool (KRINGS *et al.* 1997; NOONAN *et al.* 2006; SERRE *et al.* 2004). However, indirect studies of patterns of linkage disequilibrium (LD) in contemporary human populations have consistently found support for admixture between 'archaic' human groups (such as Neanderthals) and modern humans (GARRIGAN *et al.* 2005a; GARRIGAN *et al.* 2005b; HAMMER *et al.* 2011; LACHANCE *et al.* 2012; PLAGNOL and WALL 2006; WALL *et al.* 2009).

A detailed analysis of a draft Neanderthal genome and five low-coverage (4X) human sequences estimated that Neanderthals made a 1 – 4 % contribution to the gene pool of modern non-African populations (GREEN *et al.* 2010). The presence of 'Neanderthal DNA' in East Asians and Melanesians was initially surprising because the archaeological record shows that Neanderthals and early modern humans coexisted

only in Europe and western Asia. Green and colleagues hypothesized that Neanderthals and modern humans came into contact and interbred in the Middle East roughly 50 – 80 Kya, prior to the divergence of modern day European and Asian populations.

GREEN *et al.* (2010) presented three kinds of evidence in favor of interbreeding. First, they found (using D-statistics, a new measure of genetic similarity introduced in that paper) that the three sampled non-African genome sequences (from a French, a Han Chinese, and a Papua New Guinean) are more similar to the Neanderthal sequence than is either of the two sampled African sequences (from a San and a Yoruban). Second, they identified several haplotypes that are in low frequency in Europeans, absent from Africans, and present in the Neanderthal sequence, which suggests those haplotypes were derived from Neanderthals. Third, they found many more genomic fragments in a European genome than in an African genome that have low divergence to the Neanderthal genome.

Admixture between modern humans and Neanderthals within the past 100 Kyr is only one possible explanation for these D-statistic patterns. Green et al. noted that another potential explanation is ancient population subdivision within Africa before both Neanderthals and modern humans left Africa (cf. GREEN *et al.* 2010, Fig. 6).If there had been long-lived (e.g., > 500 Kyr) population structure within Africa, and both Neanderthals and non-African AMH came from the same 'source' subpopulation, then Neanderthals would be more similar to non-Africans in the absence of any recent admixture between AMH and Neanderthals (see Figure 1a). This intuitive argument was confirmed by the simulation studies of DURAND *et al.* (2011) and ERIKSSON and MANICA

(2012), but these studies did not account for the other two lines of evidence summarized above. Two other studies have shown that the ancient-subdivision model is incompatible with other aspects of the data. YANG *et al.* (2012) demonstrated that recent admixture (Figure 1b) could be distinguished from ancient subdivision (Fig. 1a) by computing the frequency spectrum of modern humans, conditioned on the Neanderthal sequence having the derived allele and an African sequence having the ancestral allele. This double conditioning enriches for alleles introduced by recent admixture if it occurred. Yang and colleagues found that the doubly conditioned frequency spectrum in Europeans and in East Asians is consistent with recent admixture, not with ancient subdivision. Separately, an analysis of the extent of LD at closely linked sites also concluded that the data were consistent with recent admixture and not with ancient subdivision (SANKARARAMAN *et al.* 2012).

In this study, we revisit the question of Neanderthal admixture using an expanded data set of 42 high-coverage (>45X) modern human genomic sequences and we take advantage of the recent high-coverage Denisova genome (MEYER *et al.* 2012) to obtain more refined estimates of admixture proportions. We use two complementary methods of analysis. One is the D-statistic method introduced by GREEN *et al.* (2010). D-statistics reflect site-by-site differences. Because we have multiple individuals from each of several populations we can quantify the extent of variation in D-statistics among pairs of individuals from the same two populations and obtain greater statistical power by combining estimates among all pairs. The second method is an LD-based method similar to one introduced by WALL (2000) and PLAGNOL and WALL (2006) for identifying putatively introgressed regions in modern human genomes. We use the draft

Neanderthal genome to identify segments in the modern human genome that were derived from admixture with Neanderthals. This method is similar to the one used by GREEN *et al.* (2010) but is less restrictive and allows quantification of the differences in the number of admixed segments in different populations.

Using both of these methods, we show there was more Neanderthal admixture into East Asian populations than into European populations. This conclusion is consistent with that of MEYER *et al.* (2012), which was based on the analysis of a smaller number of modern human sequences. By using the high coverage Denisova genome, we are able to show that the admixture rate into East Asians is 40% higher than into Europeans. We conclude that admixture between Neanderthals and modern humans did not occur at a single time and place, as suggested by GREEN *et al.* (2010). Some of it had to have occurred after the separation of East Asians and Europeans. Further, we show that there was significant Neanderthal admixture into the Maasai population of East Africa, probably because of secondary contact with a non-African population rather than admixture directly from Neanderthals.

MATERIALS AND METHODS

Complete Genomics data: We downloaded data from 69 publicly available genome sequences from the Complete Genomics website (http://www.completegenomics.com/public-data/). Complete Genomics sequenced a Yoruba (YRI) trio, a CEPH/Utah (CEU) pedigree family of 17 family members, a Puerto Rican (PUR) trio, and a diversity panel from ten different populations. Combining these

data sets and using only non-related, non-admixed individuals, we have a sample size of 42 individuals representing nine different populations (Table 1). In addition to 36 members of the diversity panel, we also used the parents from the YRI trio, and the maternal and paternal grandparents in the CEU pedigree. The individual genomes were sequenced to a minimum 45-fold coverage (DRMANAC *et al.* 2010). The eight populations are Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Han Chinese from Beijing, China (CHB), Gujarati Indians from Houston, Texas, USA (GIH), Japanese from Tokyo, Japan (JPT), Luhya from Webuye, Kenya (LWK), Maasai from Kinyawa, Kenya (MKK), Tuscans from Italy (TSI), and Yoruba from Ibadan, Nigeria (YRI). Samples from three other populations were also available from Complete Genomics, those of Mexican ancestry in Los Angeles, CA (MXL), African Americans from Southwest Arizona (ASW), and the Puerto Ricans from Puerto Rico (PUR), but these were excluded from our analysis because of recent intercontinental admixture. All genomic data were downloaded from Complete Genomics' ftp site (ftp://ftp2.completegenomics.com/). We used two separate pipelines for filtering and processing the data, optimized for the different analyses performed (see below).

D-statistic filtering: For the D-statistic analyses, each individual genome was aligned with the human genome assembly hg19 for consistency with the available assembly of the Neanderthal genome. Since our results were somewhat unexpected, we prepared the data for analysis in two different ways to check for consistency. We denote these Analysis A and Analysis B.

For Analysis A, we used the release of the file format version 2.0 (software version 2.0.0.26) that was generated September 2011. This version was mapped to the human reference genome hg19. We also downloaded the chimpanzee genome pantro2 aligned to hg19 from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsPanTro2/). The Neanderthal sequence was obtained by pooling reads from the three Vindija bones (SL Vi33.16, SL Vi33.25, and SL Vi33.26) that were aligned to the reference human genome (GREEN *et al.* 2010). The Neanderthal data were downloaded from the UCSC genome browser (http://genome.ucsc.edu/Neandertal/). To match the filtering used in the original GREEN *et al.* (2010) study, we used only sites with a mapping quality score (MAPQ) of at least 90 and a sequence quality higher than 40. On average, the coverage of the Neanderthal genome was about 1.3 fold. We kept only sites that had one, two, or three reads.

After filtering out any insertions, deletions, or ambiguously called sites in the Complete Genomics data, we merged them with the chimpanzee and Neanderthal genomes. We kept only sites that had no more than two alleles in any of the human genomes and at which alleles were called for each human, the chimp, and the Neanderthal. Furthermore, we considered only transversion differences.

For Analysis B, we re-downloaded the genomic data from the Complete Genomics website (ftp://ftp2.completegenomics.com/, software version 2.0.2.15, file format version 2.0, February 2012) These sequences were aligned to hg18. We applied a less stringent filter of the Neanderthal data: the filtering for mapping quality and sequence quality remained the same as in Analysis A, but there were no restrictions on the number of reads per site. Finally, instead of considering the chimp genome as the

outgroup, we used the ancestral alleles defined by the 1000 Genomes Project from the EPO (Enredo-Pecan-Ortheus) pipeline (PATEN *et al.* 2008a; PATEN *et al.* 2008b); data downloaded from ftp://ftp.1000genomes.ebi.ac.uk/). We refer to this outgroup as the reconstructed common ancestor (RCA).

For samples from any two populations compared, we filtered out any insertions, deletions, or ambiguously called sites. These genomic samples were then merged with the Neanderthal genome and the RCA outgroup. This differs from Analysis A, where all populations were merged with the Neanderthal and chimp genome prior to any comparisons between populations. We only considered sites where the difference between the ancestral allele from the RCA and the alternate allele is a transversion, as we did in Analysis A.

We also obtained the high coverage Denisova genome from MEYER *et al.* (2012). The genome was aligned to the human reference genome (hg19) and the average coverage was about 30X. We filtered out all sites that had less than 16 reads or more than 46 reads. We merged these data with the data from Analysis A to compute the f-statistic.

LD-based analysis filters: Since the LD-based analyses primarily utilize patterns of extant genetic variation (and only secondarily use the draft Neanderthal genome), we aligned variant calls to the updated human genome assembly (hg19), included both transitions and transversions, and imposed more stringent filters to throw out repetitive regions. Specifically, a custom series of Perl/C scripts and cgatools v1.3.0.9 were used to get a common set of variants from each individual. Using the CGI's variant file, all polymorphic regions containing SNPs were identified and reconstructed according to

CGI's descriptions. These regions were then filtered for SNPs in such a way that both alleles were known for a given individual and were not part of a complex variant (for example: a SNP on one haploid phase and a deletion on the other phase). We then pooled all unique SNP positions from the full panel of samples and removed all SNPs located within repeats and segmental duplications with a minimum size of 50bps. Structural variants (dgv track on UCSC), self chain (identity < 90%, UCSC self chain track), segmental duplications (UCSC), microsatellites (UCSC), simple tandem repeats (UCSC) and repeat masked sequence (UCSC) were also excluded. The final list of SNPs were then used by CGI's "snpdiff" tool to extract each sample's base calls relative to the human reference genome (hg19, Build 37). The "snpdiff" output was then reformatted to ms, PLINK and other text based formats for further analyses.

Subsequently, we identified numerous regions where all/most individuals had heterozygous SNP calls but only one homozygous genotype was present. These regions likely reflect either alignment errors due to the Complete Genomics short-read sequencing technology or errors in the human reference genome sequence. We excluded all regions that included sites where over half of the individuals are heterozygous and only one homozygous genotype is present. The coordinates for these regions are available from the authors upon request.

Denisova sequence reads (REICH *et al.* 2010), mapped to the human reference genome hg18, were downloaded from the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg18&c=chrX&g=bamSLDenisova). Consensus Neanderthal sequence generated from three bones and aligned to the human reference genome hg18 was downloaded from the Ensembl genome browser

(http://neandertal.ensemblgenomes.org/data_info.html). Samtools 0.1.18 (LI *et al.* 2009) was used to convert the BAM files into a pileup alignment (mpileup arguments: -B -q5 -Q30) of each ancient hominin genome and hg18 for the region of interest. To compare modern human sequence tracks to ancient hominid sequences, hg19 coordinates of interest were converted to hg18 coordinates using the UCSC genome browser tool liftOver and extracted from the pileup alignments via custom perl scripts. To further compare the human sequences to sequences of other primate genomes, another custom perl script was used to extract the same hg19 coordinates of interest from a subset of the genomes in the UCSC MultiZ alignments found at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/. Computations were performed using the UCSF Biostatistics High Performance Computing System.

D-statistics and estimates of admixture rates: D-statistics, introduced by (GREEN *et al.* 2010), are summary statistics for genome sequences from four populations. Two populations, P_1 and P_2 are compared to a test population, P_3. The fourth population P_4 is used as an outgroup to determine which allele is ancestral at each site. In our case, P_4 is the chimpanzee reference sequence (pantro2) denoted by *C*, and P_3 is the Neanderthal sequence, denoted by *N*. P_1 and P_2 are two human sequences. The chimp reference sequence is assumed to have the ancestral allele, denoted by *A*. *D* is computed only for sites at which both the Neanderthal and one but not both of the human sequences have a different allele, assumed to be derived and denoted by *B*. That is, only those sites with configurations *ABBA* and *BABA* are used, where the order is P_1, P_2, P_3, P_4. The requirement that two copies of both the derived

and ancestral alleles be present greatly reduces the effect of sequencing error (DURAND *et al.* 2011).

When only a single sequence from each population is available,

$$(1) \quad D(P_1, P_2, P_3, P_4) = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

where n_{ABBA} and n_{BABA} are the numbers of sites with each of the two configurations. When diploid sequences from each individual from P_1 and P_2 are available, Then

$$(2) \quad D(P_1, P_2, P_3, P_4) = \frac{\sum_i (1 - p_i^{(1)}) p_i^{(2)} - \sum p_i^{(1)} (1 - p_i^{(2)})}{\sum_i (1 - p_i^{(1)}) p_i^{(2)} + \sum p_i^{(1)} (1 - p_i^{(2)})}$$

where $p_i^{(1)}$ and $p_i^{(2)}$ are the frequencies of the derived allele (0, 0.5, 1) in the individual in P_1 and P_2 respectively at site *i*. Equation (2) is equivalent to sampling one of the chromosomes at random from P_1 and P_2 and then using Equation (1).

GREEN et al. (2010) and DURAND et al. (2011) showed that the expected value of *D* is 0 if P_1 and P_2 form a clade and P_3 is the outgroup (Fig. 2A). These papers also showed that if there was admixture from P_3 into P_2 then E(*D)* > 0 (Fig. 1B). The magnitude of *D* depends on the admixture proportion *f*, and on the population divergence times and various effective population sizes.

REICH et al. (2010) showed that if there is a sister group of P_3, which we call P_5, that has not admixed with either P_1, P_2 or P_3 (Fig. 1C), then it is possible to estimate *f* directly. In our case, P_5 is the Denisovan genome. To estimate *f*, we define $S(P_1, P_2, P_3, P_4)$ to be the numerator of either Eq. (1) or Eq. (2). Then

(3) $$\hat{f} = \frac{S(P_1, P_2, P_5, P_4)}{S(P_1, P_3, P_5, P_4)}.$$

The intuition behind this estimator is that the denominator quantifies the excess coalescent events that occur between lineages in P_3 and P_5 because they are sister groups. Lineages in P_2 that are introduced by admixture have the same coalescent history as all lineages from P_3. Hence, the ratio is the fraction of lineages in P_2 that trace their ancestry to P_3 because of admixture (REICH *et al.* 2010). In our application of this method, we are assuming that there is no admixture from Denisovans (P_5) into the other populations (P_1, …, P_4). Although SKOGLUND and JACOBSSON (2011) have argued that there was admixture from Denisovans into East Asians, our results described below did not find evidence of this admixture for the Han Chinese and Japanese samples we analyzed. For Analysis A, we explored the variation in estimated D-statistics and admixture rates (f) for all pairs of individuals of different human populations. For Analysis B, since we did not include the Denisova genome, we estimated only D-statistics.

Randomization tests: We computed D for each pair of individuals, both within populations and between populations. We developed two randomization tests of statistical significance. Both are similar to the Mantel test. Test 1 tests whether the average D computed for one pair of populations is significantly larger than for another pair, and Test 2 tests whether the average D for a pair of populations differs significantly from 0.

For Test 1, we start with sequences from three human populations, G_1, G_2 and G_3, each containing k_1, k_2 and k_3 diploid sequences. We compute two matrices of D

values. The elements of M_1 are $D(G_{1,i}, G_{3,j}, N, C)$, where $G_{1,i}$, and $G_{3,j}$ are the i-th and j-th individuals in G_1 and G_3 (i=1,…,k_1; j=1,…,k_3). The elements of M_2 are $D(G_{2,i}, G_{3,j}, N, C)$. M_1 has k_3 rows and k_1 columns, and M_2 has k_3 rows and k_2 columns. From M_1 and M_2 the average D's are computed, D_1 and D_2. The problem is to test whether D_1=D_2. A t-test cannot be used because the elements within each matrix are not independent of each other and because the same reference population (G_3) is used to compute both matrices. Instead, we combine M_1 and M_2 into a single matrix with k_3 rows and k_1+k_2 columns. Then we randomize the columns and compute D_1 for the matrix containing the first k_1 columns and D_2 for the matrix containing the last k_2 columns. Then we compare the observed D_1–D_2 with the distribution of differences from the randomized matrices. We used a two-tailed test and used one million replicates for each test.

Test 2 is similar to Test 1, but because we compare only G_1 and G_2, a subset of one population is used in place of the reference population, G_3. For the population with the larger sample size (say G_1), we create a random partition (G_1^a, G_1^b) subject to the constraint that they differ in number by no more than one. For M_1, we compute D for all pairs of individuals in G_1^a and G_2. The elements of M_2 are $D(G_{1,i}^a, G_{1,j}^b, N, C)$ where $G_{1,i}^a$ and $G_{1,j}^b$ are the i-th and j-th individuals in the two subpopulations created by the partition. Test 1 is then applied to M_1 and M_2.

We also calculated the f-statistics for each pair of individuals. Using the same randomization tests as described above, we determined whether there were significant differences between populations in estimates of the admixture rate. Significant differences observed using the admixture rate suggest that the effect is truly due to the Neanderthal and not admixture with Denisovans.

Identifying putative archaic human regions: Previous work has shown that archaic admixture often leads to long, divergent haplotypes at low frequency (PLAGNOL and WALL 2006; WALL 2000). We define two SNPs to be 'congruent' if their diploid allele counts (i.e., 0, 1 or 2 counts of a particular allele) across individuals are completely correlated (i.e., $r^2 = 1$). We define the maximum number of pairwise congruent SNPs to be l_d, and denote the collection of rarer (MAF ≤ 0.5) alleles at each of these pairwise congruent sites to be the putative archaic haplotype. From the filtered Complete Genomics data, we then identified all regions from 8 – 100 Kb in length where $l_d \geq 30$ and $l_d / S \geq 0.1$, where S is the total number of polymorphic sites in the region. When identified regions overlapped, we took the region with the largest value of l_d / S. We also required that neighboring regions with putative archaic haplotypes congruent with each other be separated by at least 200 Kb, to avoid double-counting long archaic haplotypes. A total of 2,254 regions were identified. Of these, 411 were private to the non-African samples.

To estimate what proportion of these regions might be false positives, we simulated whole-chromosome sequence data (CHEN *et al.* 2009) under a model that incorporated both recent (intracontinental) and ancient (intercontinental) population structure (Figure 2). Specifically, we assume a panmictic ancestral population split into two daughter populations at time $T_0 = 0.6$ (using the standard coalescent scaling of $4N$ generations), with (symmetric) scaled migration rate of $M_0 = 5$. At time $T_1 = 0.05 – 0.053$, one of the ancestral populations (i.e., the 'non-African' one) experiences a population bottleneck resulting in a 100-fold reduction in population size. Then, at time $T_2 = 0.045$, each population splits into two descendant populations, connected by

migration rate M_1 = 8. While arbitrary, this model attempts to incorporate the major features of human demographic history, including intra- and intercontinental population structure and a bottleneck in the history of non-African populations, and is similar to the model used by (YANG *et al.* 2012). The results described below are qualitatively similar if other plausible values for the times and migration rates are used (Results not shown). Using N = 10,000 and an average generation time of 25 years, each unit of scaled time corresponds to a million years.

We simulated 30 different 100 Mb chromosomes using the model described above with mutation parameter $\theta = 3.5 * 10^{-4}$ / bp, recombination parameter $\rho = 4 * 10^{-4}$ / bp, and 10 individuals sampled from each of the four extant populations. The simulated number of segregating sites was substantially higher than the actual number in our filtered data. Since average l_d values are positively correlated with levels of diversity, the simulated l_d values are higher on average than expected in real data, and our choice of θ is conservative. Also, standard estimates of ρ are generally higher than the value we took (MYERS *et al.* 2005), which is also conservative for our purposes. We then tabulated the total number of regions with $l_d \geq 30$, $l_d / S \geq 0.1$, and with divergent haplotype SNPs private to the simulated 'non-African' samples. We identified a total of 3 regions that satisfied these criteria, compared with 411 regions that were identified from the actual data. This leads to an estimate of a false discovery rate of $q < 0.01$.

Identifying putative Neanderthal regions: To identify which of the 2,254 regions described above were likely to reflect recent Neanderthal admixture, we imposed the following additional criteria on the putative archaic human haplotypes:

I) The Neanderthal allele must be called at 12 or more SNPs and match the putative archaic haplotype at ≥ 70% of these SNPs

II) The Neanderthal allele and chimp allele must be called at 8 or more SNPs and the Neanderthal allele must be derived (relative to chimp) at ≥ 60% of these sites

III) The putative archaic haplotype must be at low frequency (< 5%) in the sub-Saharan African samples

The motivation for (I) is obvious, and we note that a more stringent cutoff was not used due to the poor quality of the Neanderthal genome sequence. (II) was implemented to cut down on regions that reflect shared ancestral polymorphism between modern humans and Neanderthals; it is based on an observation of (NOONAN *et al.* 2006) that recent Neanderthal admixture will lead to an increase in SNPs where Neanderthals have the derived allele. Finally, (III) reflects our prior belief that admixture with Neanderthals did not occur in Africa, and that the presence of Neanderthal alleles in Africa could only reflect more recent migration patterns. A total of 226 regions were identified that meet these additional criteria. We note in passing that the specific cutoffs used in (1) – (3) are somewhat arbitrary, but our qualitative conclusions are unchanged under a range of similar criteria (Results not shown).

We implemented a simple permutation test to assess the statistical significance of the observed difference in frequencies of Neanderthal regions in East and South Asians and Europeans. Specifically, we kept the presence/absence of Neanderthal regions for each individual constant and randomly permuted the geographic label (i.e., 'European'

vs. ‘East Asian) of the sample 100,000 times. Similar analyses were used to compare the frequency of Neanderthal regions in Maasai versus other sub-Saharan African samples.

Identifying putative Denisovan regions: Excluding the 226 Neanderthal regions identified above, we screened the remaining 2,028 putative archaic regions for Denisovan admixture using the same criteria as for Neanderthals. 30 total regions fit these criteria.

Estimating local ancestry in the Maasai: We took the filtered Complete Genomics data described at the start of this section and estimated SNP allele frequencies separately in the 13 European samples and the 13 non-Maasai African samples. These were used as proxies for the (unknown) ‘non-African’ and ‘African’ ancestral populations We then included only those SNPs with allele frequencies that differ by at least 0.3 in our analyses. We calculated the likelihood of each ancestral configuration (i.e., 0, 1 or 2 alleles inherited from the ‘non-African’ population) separately for each SNP. Then, over sliding windows of 1 Mb, we formed a composite-likelihood by multiplying together all of the single-SNP likelihoods contained in the window, and tabulated which ancestral configuration had the highest (composite) likelihood. For each SNP, we then used majority-rule to make ancestry calls using all windows containing the SNP in question. See (WALL *et al.* 2011) for further details.

RESULTS

D-statistics and estimates of *f*: The D-statistics and estimates of f we computed are summarized in Figure 3 and supplemental material Note S1, Tables S1-S9 and Figures S1-S8. Several features of the results are notable. First, we find evidence for more Neanderthal admixture into the East Asian samples than into the European samples (p = 0.001) – consistently higher D values result when East Asians are compared to one of the African populations than when Europeans are compared (Figure 3a, Table S4), and the average D is positive when East Asians are compared to Europeans (Figure 3c, Table S5). In Analysis B, comparisons with the South Asian samples are intermediate with respect to the European and East Asian samples but not in Analysis A, indicating that the South Asian sample differs from the East Asian ones but the degree of similarity to Europeans remains to be established. Also, we find evidence for a small but significant amount of Neanderthal admixture into the Maasai genomes (p ~ 0.03, Table S4). When compared to the Yoruba, the Maasai have a higher average D than the Luhya (Figure 3b, Table S4). When the Maasai are compared to all other African samples the average D is positive (Figure 3d). In addition, when East Asians and Europeans are compared to the Maasai, the average D's are somewhat lower than when they are compared to either the Yoruba or Luhya. The p-values shown in Figure 3A and 3B are from Test 1 and those in Figure 3C and 3D are from Test 2.

Tables S1-S3 show estimated values of f. The estimates of the admixture rate show that when we incorporate the Denisovan genome into our analysis, the admixture rate between East Asians and Neanderthals remains significantly higher than the admixture rate between Europeans and Neanderthals (p ~ 0.001, Table S7). The

Maasai remain significantly more genetically similar to the Neanderthals when compared to the Luhya ($p \sim 0.03$, Table S7), but the observed significant difference for the D-statistic when comparing the Maasai and Yoruba is not observed for the f-statistic ($p \sim 0.34$, Table S7), which probably reflects the lower power of using *f* as a test statistic. The admixture rates for the South Asians give same results as that for the D-statistic (Table S9).

Identifying 'Neanderthal' haplotypes: Our new method for identifying introgressed Neanderthal fragments in human populations detected 226 different putative Neanderthal regions. The relative frequencies of these putative Neanderthal haplotypes in the 42 sampled modern human individuals then provide estimates of the relative contributions of Neanderthal DNA to the gene pools of contemporary human populations. We found that on average the 'Neanderthal haplotypes' were at higher frequency in the East Asians than in the Europeans (9.6% vs. 6.4%; $p = 3.0 \times 10^{-4}$, permutation test), consistent with the *D*-statistic results presented in Figure 3. We also found evidence for a small, but statistically significant, Neanderthal contribution to the genomes of the Maasai ($p = 4.9 * 10^{-4}$), but did not find a significant difference in Neanderthal haplotype frequency between the East and South Asian samples ($p > 0.05$).

Additional test of ancient population structure: As reviewed in the introduction, there is already evidence against the hypothesis that the extra similarity of non-African populations to Neanderthals is accounted for by ancient population subdivision. To explore this point further, we took the 411 regions from our whole-genome analyses that were identified purely on the basis of their LD patterns (i.e., without using any

information from the Neanderthal genome sequence). Then, for each non-African individual, we calculated the D-statistic for those regions where the individual contained a rare, diverged haplotype. If this haplotype were recently inherited from Neanderthals, we would expect the *D* values to be strongly negative. If instead there were no recent admixture between modern humans and Neanderthals, then there is no *a priori* reason why these regions would show *D* values significantly different from 0. Recombination acting over the past 300 Kyr would break up local patterns due to shared ancestral polymorphisms to scales smaller than 0.01 cM (i.e., < 10 Kb on average). The D-values that we observe are strongly negative (average D = -0.594, compared with an average D = -0.068 for the whole genome), providing additional evidence that most of the unusual haplotypes from these 411 regions are indeed the result of recent introgression from the Neanderthal gene pool ($p << 10^{-8}$, Figure 6).

Identifying 'Denisovan haplotypes': Excluding the 226 Neanderthal regions described above, we used the same criteria to identify regions likely inherited from Denisovans. We identified a total of 30 regions, all at low frequency, and with no significant difference in frequency between populations.

Maasai admixture: Previous genetic studies have suggested that the Maasai may be an admixed population with a substantial proportion of non-African ancestry (HENN *et al.* 2011). If the non-African ancestry were due to recent (i.e., post-Neanderthal) admixture, then the observation of Neanderthal ancestry in the Maasai would not be unexpected. Alternatively, spatially explicit models of ancient population structure might explain the greater similarity between Maasai and Neanderthals relative to other sub-Saharan African groups (A. Manica, personal communication). One difference between

these alternative explanations is what they predict about the patterns of similarity across the genomes of Maasai individuals. Under a model of recent admixture, we expect Maasai genomes to show large, distinct blocks of sequence with different genetic patterns, corresponding to blocks with non-African vs. African ancestry. The average size of the non-African blocks (in Morgans) is roughly the inverse of the time (in generations) since admixture. In contrast, under a model of ancient admixture the similarity of Maasai genomes with the Neanderthal genome will be spread throughout the genome because the admixture happened much longer ago.

To distinguish between these two possibilities, we employed a composite-likelihood based approach to identifying 'African' and 'non-African' regions of ancestry across the genomes of the 4 Maasai samples (WALL *et al.* 2011). Briefly, we used the European (CEU and TSI) and other African (YRI and LWK) samples (Table 1) to estimate allele frequencies in 'non-African' and 'African' ancestral populations, and then estimated the number of alleles inherited from each ancestral population at each SNP in the genome. These extant samples may not be perfect proxies for the true ancestral populations, but the qualitative results presented below are likely to be valid.

In summary, we estimate an average of ~30% 'non-African' ancestry in each Maasai genome, and the sizes of the ancestral blocks are consistent with admixture that happened ~100 generations ago (Figure 5a). We then partitioned each Maasai genome into regions with 0, 1 or 2 inferred 'African' alleles and calculated D separately for each partition. We found that the D values are significantly more negative with increasing numbers of inferred 'non-African' alleles ($p = 2.0 * 10^{-4}$; Figure 5b). This observation

provides strong support for recent non-African gene flow into the Maasai, with the non-African alleles bringing with them low levels of Neanderthal ancestry.

DISCUSSION AND CONCLUSIONS

Our results confirm and reinforce several conclusions about admixture between Neanderthals and the ancestors of modern humans. Using a much larger number of high-coverage genome sequences than were previously analyzed for this purpose and using two complementary methods of analysis (D-statistics and detection of introgressed Neanderthal segments), we confirm the conclusion of MEYER *et al.* (2012) that East Asians (Han Chinese and Japanese) are more similar to the published Neanderthal sequence than are Europeans. Because we have analyzed more modern human sequences than MEYER *et al.* (2012) did, we are able to show the extent of variation within both Asian and African populations. We also confirm the conclusions of YANG *et al.* (2012) and SANKARARAMAN *et al.* (2012) that the similarity of both Europeans and East Asians to Neanderthals is the result of recent admixture and not ancient population subdivision. Finally, we used the high-coverage Denisova sequence of MEYER *et al.* (2012) to determine that the admixture rate (f) into East Asians is roughly 40% higher than into Europeans.

We were not able to confirm the conclusion of SKOGLUND and JAKOBSSON (2011) that there was Denisovan admixture into East Asians. We did not detect any difference in the number of apparent Denisovan segments in Europeans and East Asians. The

East Asian genomes were analyzed, however, were from northern East Asia (Beijing and Tokyo), not from southern East Asia where Skoglund and Jakobsson found the strongest signal of admixture with Denisovans.

Our results and those of MEYER *et al.* (2012) imply that the relatively simple admixture scenario proposed by (GREEN *et al.* 2010) needs to be altered. At least two separate episodes of admixture between Neanderthals and modern humans must have occurred, and at least one of those episodes must have occurred after the separation of the ancestors of modern Europeans and East Asians. Rather than have two distinct episodes of admixture, it seems more plausible that admixture took place over a protracted period 50-80 Kya. During that period the ancestors of Europeans diverged and subsequently experienced less admixture than the ancestors of East Asians. This scenario is consistent with the simulation models of CURRAT and EXCOFFIER (2011) and SKOGLUND and JAKOBSSON. (2011)

If this scenario is correct, the time of separation of the ancestors of modern European and East Asian populations is constrained. Since there is no archeological record of Neanderthals in the past ~30 thousand years, it follows that the separation of Europeans from East Asians had to have occurred before Neanderthals went extinct. Consequently, estimates of East Asian-European population divergence of less than 30 thousand years ago (GRAVEL *et al.* 2011; GUTENKUNST *et al.* 2009) are unlikely to be correct. This timeframe is also supported by a 40 – 50 Kya modern human fossil recently found in China (FU *et al.* 2013).

Our two analyses yielded slightly different results for the Gujarati (South Asian) samples. However, it would not be surprising if the true level of Neanderthal ancestry in

South Asians was intermediate between Europeans and East Asians because previous studies have shown gradients in genetic ancestry across Eurasia (ROSENBERG *et al.* 2002).

Our finding of Neanderthal admixture into the Maasai was initially surprising, given the lack of evidence that Neanderthals ever crossed into Africa or that the ancestors of the Maasai were ever in the Middle East. Although direct contact between the two groups in the past is theoretically possible, our results are more consistent with a scenario involving recent admixture between the ancestors of the Maasai and one or more (historically) non-African groups with Neanderthal ancestry several thousand years ago. This interpretation is broadly consistent with recent findings of African admixture into Middle Eastern and Southern European populations during the same timescale (MOORJANI *et al.* 2011), and a greater genetic similarity between East African and non-African samples than between West African and non-African samples (TISHKOFF *et al.* 2009). Together these studies provide additional support for the hypothesis that admixture between genetically diverged groups is a common feature of human demographic history.

The new picture of human and Neanderthal ancestry that emerges from our results is almost certainly not complete, and our results suggest that intracontinental variation in levels of Neanderthal ancestry may be common. With the current rate of progress in whole genome sequencing and the possibility of additional draft genomes from specimens of archaic individuals, we will soon learn more about the admixture process. In particular, the construction of 'archaic admixture maps' detailing the distribution of archaic DNA segments in different modern human populations will help us to infer the

timing, locations, and exact numbers of introgression events and the role that archaic admixture may have played in the evolution of the AMH genome.

Acknowledgments This work was supported in part by NIH grants R01-GM40282 (to M. S.), R01-HG005226 (to J. D. W. and M. F. H.), and T32 HG 00047 (Training grant), as well as NSF GRFP DGE 1106400 (to M. A. Y.).

REFERENCES

CHEN, G. K., P. MARJORAM and J. D. WALL, 2009 Fast and flexible simulation of DNA sequence data. Genome Res **19:** 136-142.

CURRAT, M., and L. EXCOFFIER, 2011 Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc Natl Acad Sci U S A **108:** 15129-15134.

DRMANAC, R., A. B. SPARKS, M. J. CALLOW, A. L. HALPERN, N. L. BURNS *et al.*, 2010 Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science **327:** 78-81.

DUARTE, C., J. MAURICIO, P. B. PETTITT, P. SOUTO, E. TRINKAUS *et al.*, 1999 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.

DURAND, E. Y., N. PATTERSON, D. REICH and M. SLATKIN, 2011 Testing for ancient admixture between closely related populations. Mol Biol Evol **28:** 2239-2252.

ERIKSSON, A., and A. MANICA, 2012 Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci U S A **109:** 13956-13960.

FINLAYSON, C., 2004 *Neanderthals and Modern Humans: An Ecological and Evolutionary perspective*. Cambridge University Press, Cambridge, UK.

FU, Q., M. MEYER, X. GAO, U. STENZEL, H. A. BURBANO *et al.*, 2013 DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A, in press.

GARRIGAN, D., Z. MOBASHER, S. B. KINGAN, J. A. WILDER and M. F. HAMMER, 2005a Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics **170:** 1849-1856.

GARRIGAN, D., Z. MOBASHER, T. SEVERSON, J. A. WILDER and M. F. HAMMER, 2005b Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol **22:** 189-192.

GRAVEL, S., B. M. HENN, R. N. GUTENKUNST, A. R. INDAP, G. T. MARTH *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A **108:** 11983-11988.

GREEN, R. E., J. KRAUSE, A. W. BRIGGS, T. MARICIC, U. STENZEL *et al.*, 2010 A Draft Sequence of the Neandertal Genome. Science (Washington D C) **328:** 710-722.

GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet **5:** e1000695.

HAMMER, M. F., A. E. WOERNER, F. L. MENDEZ, J. C. WATKINS and J. D. WALL, 2011 Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A **108:** 15123-15128.

HENN, B. M., C. R. GIGNOUX, M. JOBIN, J. M. GRANKA, J. M. MACPHERSON *et al.*, 2011 Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A **108:** 5154-5162.

HUBLIN, J. J., 2009 Out of Africa: modern human origins special feature: the origin of Neandertals. Proc Natl Acad Sci U S A **106:** 16022-16027.

KRINGS, M., A. STONE, R. W. SCHMITZ, H. KRAINITZKI, M. STONEKING *et al.*, 1997 Neandertal DNA sequences and the origin of modern humans. Cell **90:** 19-30.

LACHANCE, J., B. VERNOT, C. C. ELBERS, B. FERWERDA, A. FROMENT *et al.*, 2012 Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell **150:** 457-469.

LAHR, M. M., 1994 The Multiregional Model of Modern Human Origins - a Reassessment of Its Morphological Basis. J Hum Evol **26:** 23-56.

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078-2079.

MEYER, M., M. KIRCHER, M. T. GANSAUGE, H. LI, F. RACIMO *et al.*, 2012 A high-coverage genome sequence from an archaic Denisovan individual. Science **338:** 222-226.

MOORJANI, P., N. PATTERSON, J. N. HIRSCHHORN, A. KEINAN, L. HAO *et al.*, 2011 The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet **7:** e1001373.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

NOONAN, J. P., G. COOP, S. KUDARAVALLI, D. SMITH, J. KRAUSE *et al.*, 2006 Sequencing and analysis of Neanderthal genomic DNA. Science **314:** 1113-1118.

PATEN, B., J. HERRERO, K. BEAL, S. FITZGERALD and E. BIRNEY, 2008a Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res **18:** 1814-1828.

PATEN, B., J. HERRERO, S. FITZGERALD, K. BEAL, P. FLICEK *et al.*, 2008b Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res **18:** 1829-1843.

PLAGNOL, V., and J. D. WALL, 2006 Possible ancestral structure in human populations. PLoS Genet **2:** e105.

REICH, D., R. E. GREEN, M. KIRCHER, J. KRAUSE, N. PATTERSON *et al.*, 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468:** 1053-1060.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD *et al.*, 2002 Genetic structure of human populations. Science **298:** 2381-2385.

SANKARARAMAN, S., N. PATTERSON, H. LI, S. PAABO and D. REICH, 2012 The Date of Interbreeding between Neandertals and Modern Humans. PLoS Genet **8:** e1002947.

SERRE, D., A. LANGANEY, M. CHECH, M. TESCHLER-NICOLA, M. PAUNOVIC *et al.*, 2004 No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol **2:** E57.

SKOGLUND, P., and M. JAKOBSSON, 2011 Archaic human ancestry in East Asia. Proc Natl Acad Sci U S A **108:** 18301-18306.

STRINGER, C. B., and P. ANDREWS, 1988 Genetic and fossil evidence for the origin of modern humans. Science **239:** 1263-1268.

STRINGER, C. B., and J. HUBLIN, 1999 New age estimates for the Swanscombe hominid, and their significance for human evolution. J Hum Evol **37:** 873-877.

TISHKOFF, S. A., F. A. REED, F. R. FRIEDLAENDER, C. EHRET, A. RANCIARO *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science **324:** 1035-1044.

TRINKAUS, E., 2007 European early modern humans and the fate of the Neandertals. Proc Natl Acad Sci U S A **104:** 7367-7372.

WALL, J. D., 2000 Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154:** 1271-1279.

WALL, J. D., R. JIANG, C. GIGNOUX, G. K. CHEN, C. ENG *et al.*, 2011 Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. Mol Biol Evol **28:** 2231-2237.

WALL, J. D., K. E. LOHMUELLER and V. PLAGNOL, 2009 Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol Biol Evol **26:** 1823-1827.

YANG, M. A., A. S. MALASPINAS, E. Y. DURAND and M. SLATKIN, 2012 Ancient structure in Africa unlikely to explain neanderthal and non-african genetic similarity. Mol Biol Evol **29:** 2987-2995.

Table 1. 42 individual genome sequences from Complete Genomics included in our study

ID	Population	ID	Population
NA06985	CEU	NA21732	MKK
NA06994	CEU	NA21733	MKK
NA07357	CEU	NA21737	MKK
NA10851	CEU	NA21767	MKK
NA12004	CEU	NA18940	JPT
NA12889	CEU	NA18942	JPT
NA12890	CEU	NA18947	JPT
NA12891	CEU	NA18956	JPT
NA12892	CEU	NA20502	TSI
NA18526	CHB	NA20509	TSI
NA18537	CHB	NA20510	TSI
NA18555	CHB	NA20511	TSI
NA18558	CHB	NA18501	YRI
NA20845	GIH	NA18502	YRI
NA20846	GIH	NA18504	YRI
NA20847	GIH	NA18505	YRI
NA20850	GIH	NA18508	YRI
NA19017	LWK	NA18517	YRI
NA19020	LWK	NA19129	YRI
NA19025	LWK	NA19238	YRI
NA19026	LWK	NA19239	YRI

Figure Legends

Figure 1. Simplified versions of models of ancient population structure (**A**) or recent admixture (**B**) that can explain the observed levels of divergence between modern human genomes and the draft Neanderthal genome.

Figure 2. Schematic of a model of recent and ancient population structure without admixture used in simulations. See text for details.

Figure 3. Summary of significance tests for average values of *D*. Positive values indicate that the second sequence is more similar to the Neanderthal genome than the first sequence. In all parts, the box plots indicate the range of *D* values obtained for pairs of individuals from the populations indicated. Parts A and B are box plots of individual *D* statistics computed for each individual from the specified population compared with each Yoruban. The *p* values are from the randomization test, Test 1, of significant differences in the average *D* values for different pairs of populations. Parts C and D show box plots of individual *D* statistics computed for every pair of individuals in the specified populations. The *p* values are from the randomization test, Test 2, of significant differences of the average *D* from 0. See also Table 2.

Figure 4. Distribution of the number of putative Neanderthal regions for each Eurasian individual. European genomes are colored in green, East Asian genomes are colored in red and South Asian genomes are colored in black.

Figure 5. Recent and ancient admixture in the Maasai. A) A representative plot of number of estimated 'African' alleles across the first 30 Mb of chromosome 1 in one of the Maasai genomes. B) Estimated values of D for portions of the genome estimated to contain 0, 1 or 2 'non-African' alleles.

Figure 6. Box plot showing the average D across the whole genomes of the non-African individuals compared with the average D (for the same individuals) across regions identified as having unusual patterns of LD (i.e., putative archaic regions).

Figure 1

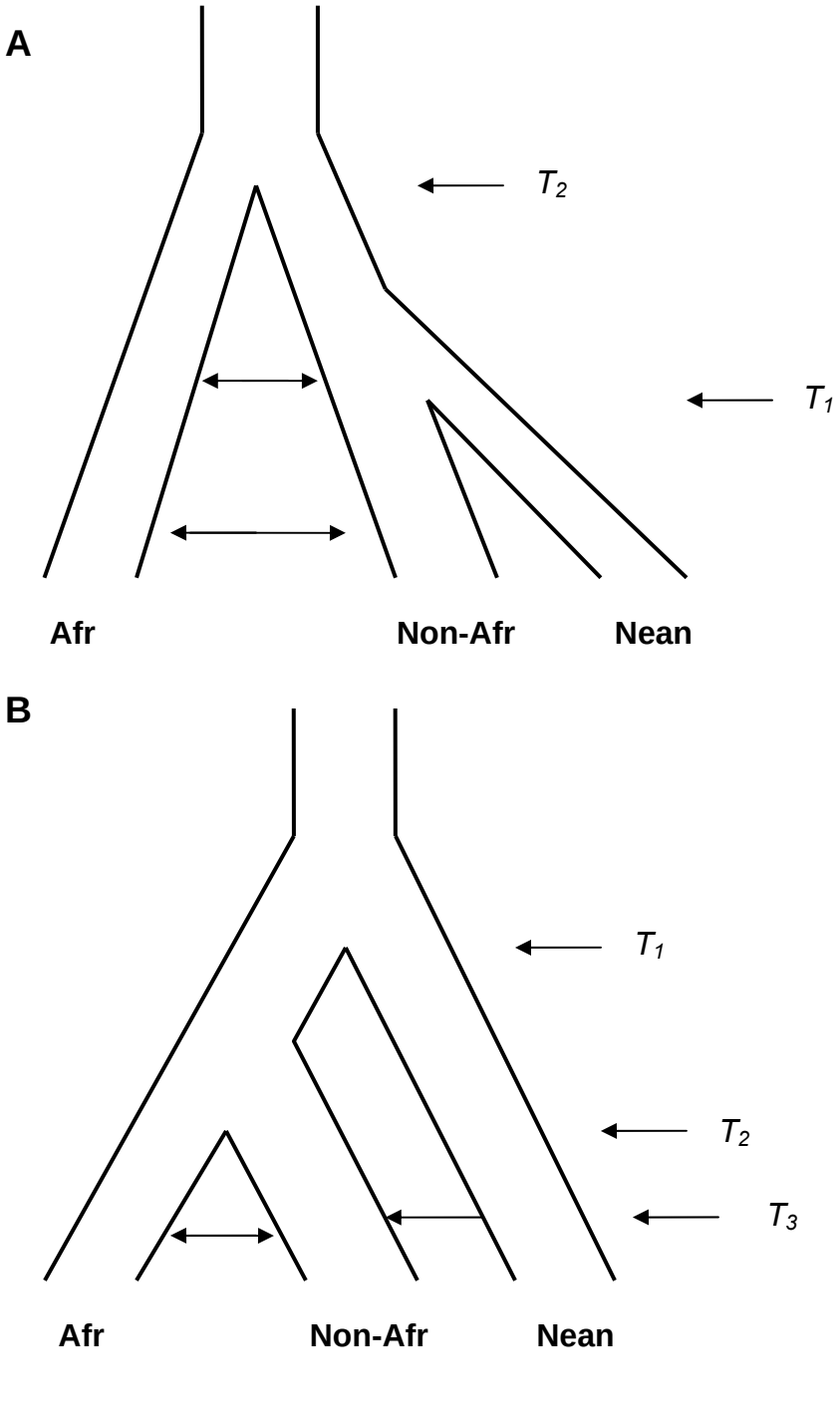
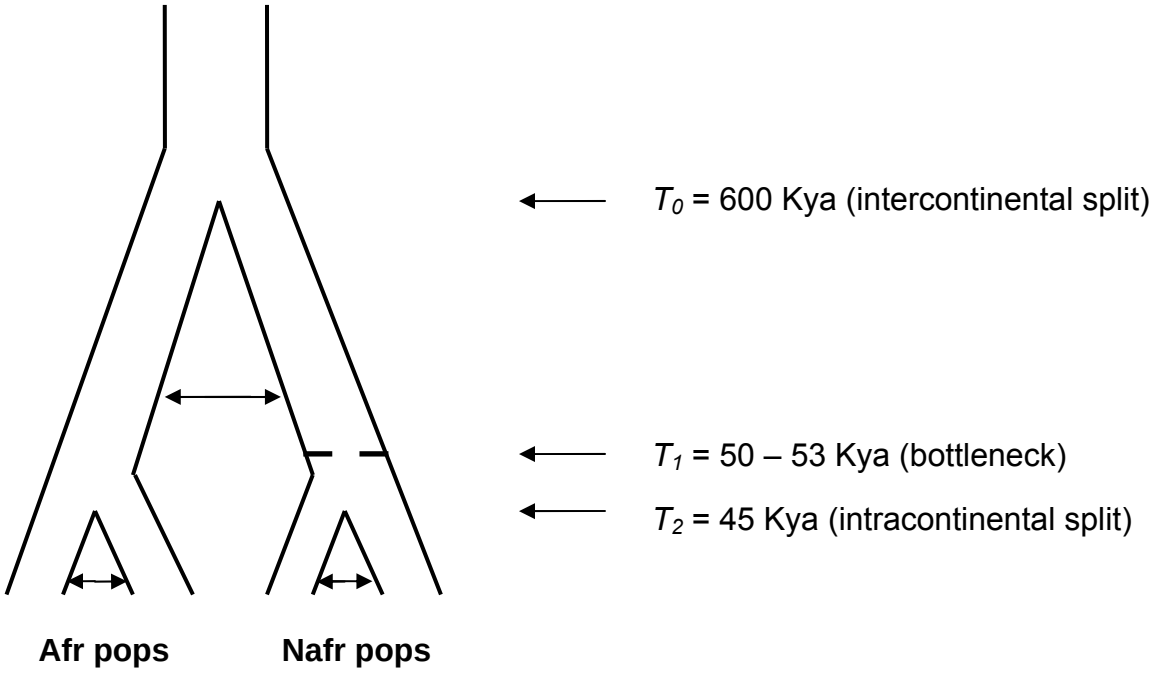


Figure 2


T_0 = 600 Kya (intercontinental split)
T_1 = 50 – 53 Kya (bottleneck)
T_2 = 45 Kya (intracontinental split)
Afr pops
Nafr pops

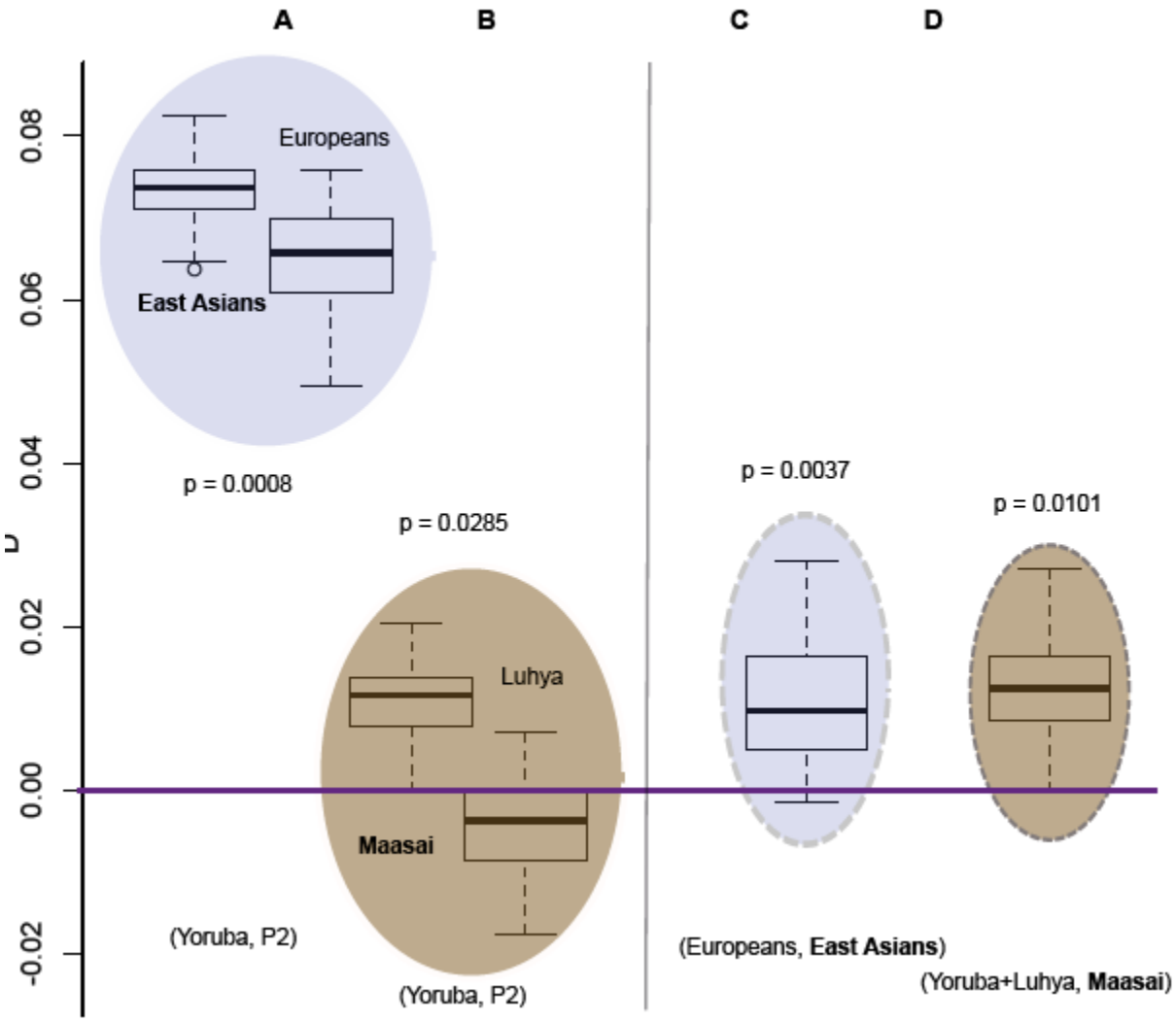

A
B
C
D
0.08
0.06
0.04
0.02
0.00
-0.02
Europeans
East Asians
p = 0.0008
p = 0.0285
Luhya
Maasai
(Yoruba, P2)
(Yoruba, P2)
p = 0.0037
p = 0.0101
(Europeans, East Asians)
(Yoruba+Luhya, Maasai)

Figure 3.

Figure 4.

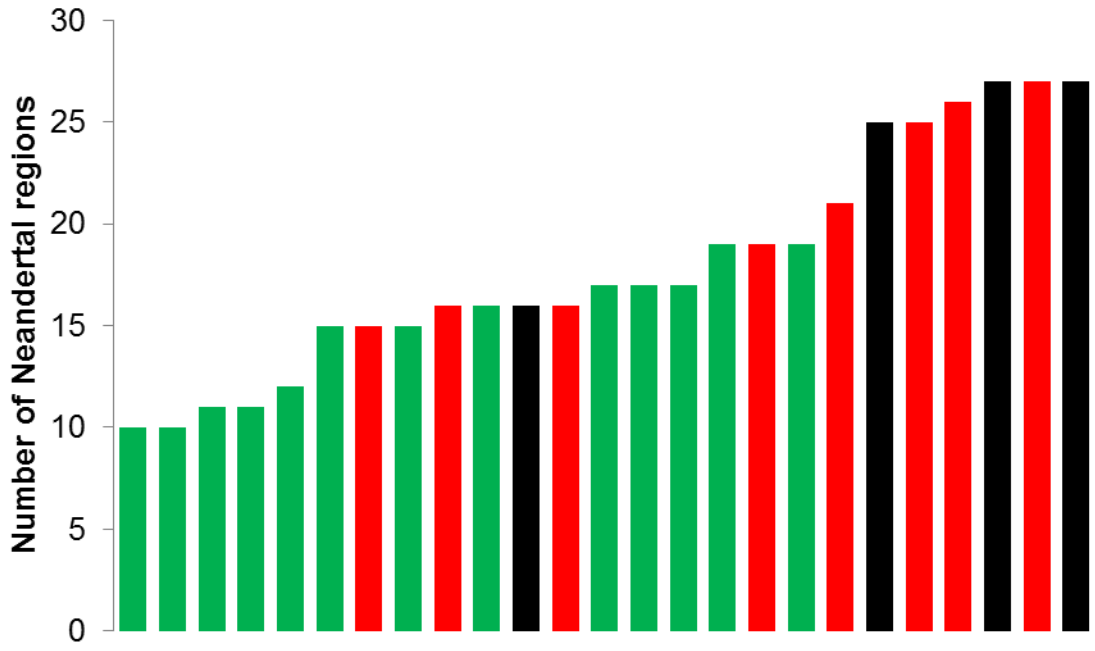
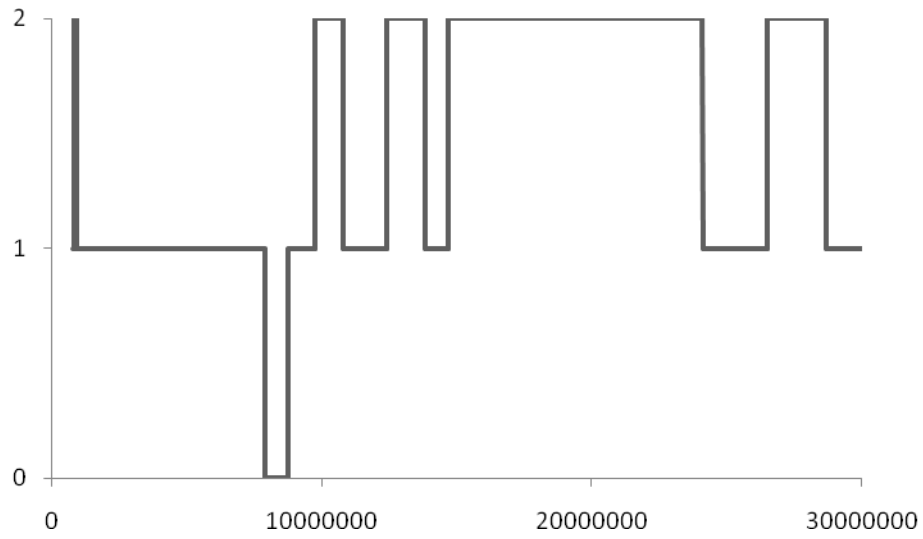


Figure 5

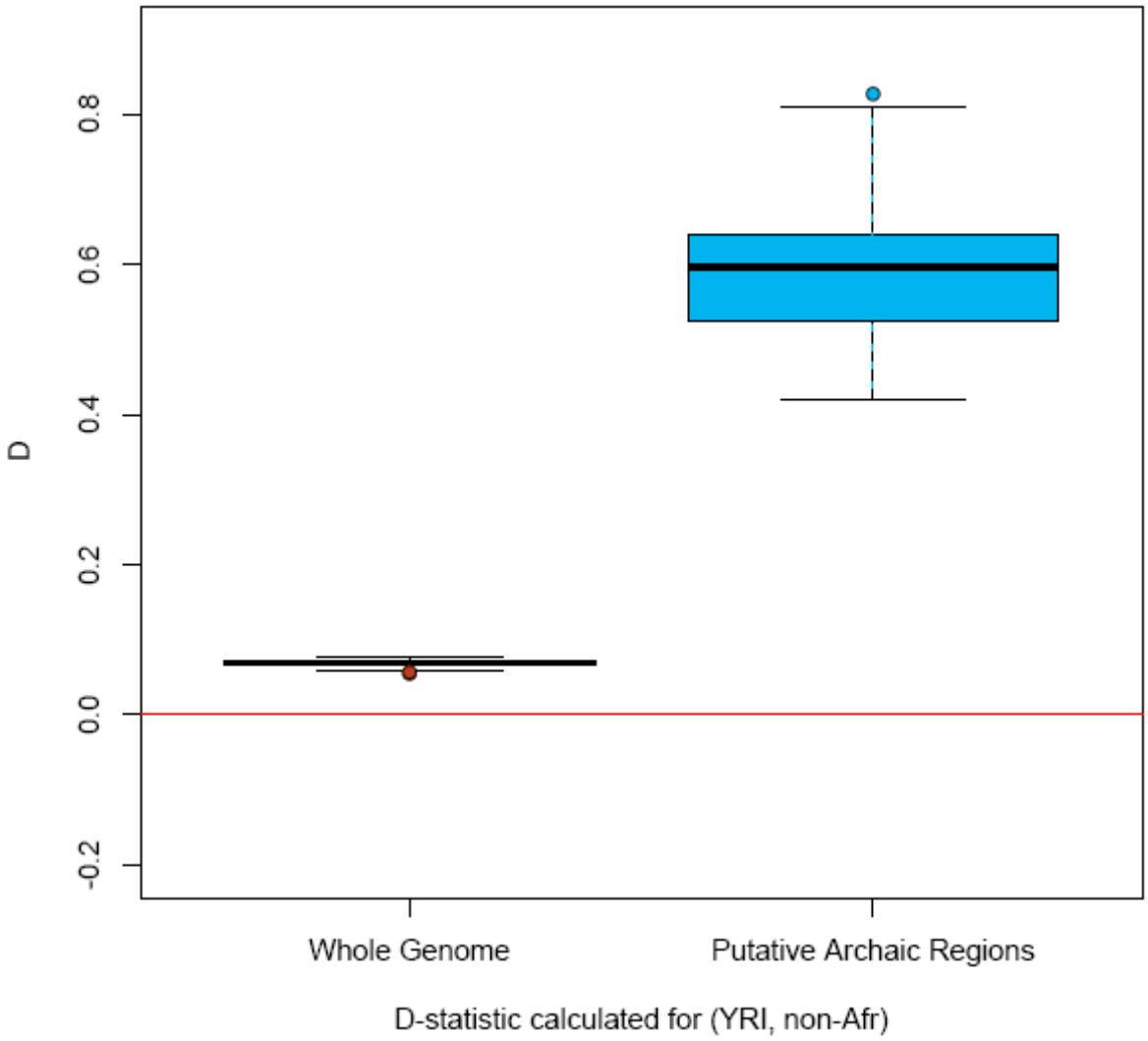
A


2
1
0
0
10000000
20000000
30000000

B


0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06

Figure 6


D
0.8
0.6
0.4
0.2
0.0
-0.2
Whole Genome
Putative Archaic Regions
D-statistic calculated for (YRI, non-Afr)