Published in final edited form as:

Nature. 2014 March 20; 507(7492): 354–357. doi:10.1038/nature12961.

The landscape of Neandertal ancestry in present-day humans

Sriram Sankararaman[1,2], **Swapan Mallick**[1,2], **Michael Dannemann**[3], **Kay Prüfer**[3], **Janet Kelso**[3], **Svante Pääbo**[3], **Nick Patterson**[1,2], and **David Reich**[1,2,4]

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA 02115

[2]Broad Institute of Harvard and MIT, Cambridge, MA, USA 02142

[3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany 04103

[4]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA 02115

Abstract

Analyses of Neandertal genomes have revealed that Neandertals have contributed genetic variants to modern humans[1–2]. The antiquity of Neandertal gene flow into modern humans means that regions that derive from Neandertals in any one human today are usually less than a hundred kilobases in size. However, Neandertal haplotypes are also distinctive enough that several studies have been able to detect Neandertal ancestry at specific loci[1,3–8]. Here, we have systematically inferred Neandertal haplotypes in the genomes of 1,004 present-day humans[12]. Regions that harbor a high frequency of Neandertal alleles in modern humans are enriched for genes affecting keratin filaments suggesting that Neandertal alleles may have helped modern humans adapt to non-African environments. Neandertal alleles also continue to shape human biology, as we identify multiple Neandertal-derived alleles that confer risk for disease. We also identify regions of millions of base pairs that are nearly devoid of Neandertal ancestry and enriched in genes, implying selection to remove genetic material derived from Neandertals. Neandertal ancestry is significantly reduced in genes specifically expressed in testis, and there is an approximately 5-fold reduction of Neandertal ancestry on chromosome X, which is known to harbor a disproportionate fraction of male hybrid sterility genes[20–22]. These results suggest that part of the reduction in Neandertal ancestry near genes is due to Neandertal alleles that reduced fertility in males when moved to a modern human genetic background.

To search systematically for Neandertal haplotypes, we developed a method based on a Conditional Random Field[9] (CRF) that combines information from three features of genetic variation that are signatures of Neandertal ancestry (SI 1; Extended Data Fig. 1). The first is

To whom correspondence should be addressed: sankararaman@genetics.med.harvard.edu or reich@genetics.med.harvard.edu.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. The tiling path of confidently inferred Neandertal haplotypes, as well as the Neandertal introgression map, can be found at http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html.

Author contributions

SS, NP, SP and DR conceived of the study. SS, SM MD, KP, JK and DR performed analyses. JK, SP, NP and DR supervised the study. SS and DR wrote the manuscript with help from all coauthors.

The authors declare no competing financial interests.

the allelic pattern at a single nucleotide polymorphism (SNP): if a non-African carries a derived allele seen in Neandertals but absent from the West African Yoruba (YRI), it likely originates from Neandertals. The second is high sequence divergence to all YRI haplotypes but low divergence to Neandertal. The third is a length consistent with interbreeding 37–86 thousand years ago[10]. We trained the CRF using simulations[11], and established its robustness to deviations from the assumed demography (SI 2).

We screened for Neandertal haplotypes in the 1000 Genomes Project Phase 1 data [12] (1KG) using the Altai Neandertal genome of 52-fold average coverage to determine alleles present in Neandertals[2], a 6-primate consensus to determine ancestral alleles[13], and 176 YRI genomes as a reference panel assumed to harbor no Neandertal ancestry (Fig. 1a). Table 1 reports the mean and standard deviation across individuals of the fraction of their ancestry confidently inferred to be Neandertal (probability >90%). Fig. 1b and Extended Data Fig. 2 plot the fraction of European (n=758) and East Asian (n=572) haplotypes that descend from Neandertals at each genomic location (SI 3). We created a tiling path of inferred Neandertal haplotypes that spans 1.1 Gigabases (Gb) over 4,437 contigs (SI 4), thus filling in gaps in the Neandertal sequence over a number of repetitive regions that cannot be reconstructed from short ancient DNA fragments (Extended Data Fig. 3).

Four features of the Neandertal introgression map suggest that it is producing reasonable results. First, when we infer Neandertal ancestry using low-coverage data from Croatian Neandertals[1] we obtain correlated inferences (Spearman rank correlation $\rho_{\text{Spearman}}=0.88$ in Europeans; SI 3). Second, in the African Luhya (LWK), the proportion of the genome inferred to be Neandertal is 0.08%, an order of magnitude smaller than in non-Africans (Table 1). Third, the proportion of confidently inferred Neandertal ancestry has a mean of 1.38% in East Asians and 1.15% in Europeans (Table 1), consistent with previous reports of more Neandertal ancestry in East Asians than in Europeans [7,14]. Fourth, the standard deviation in Neandertal ancestry within populations is 0.06–0.10%, in line with theoretical expectation (SI 3) and showing that Neandertal ancestry calculators that estimate differences on the order of a percent [15] are largely inferring noise.

The Neandertal introgression map reveals locations where Neandertal ancestry is inferred to be as high as 62% in East Asians and 64% in Europeans (Fig. 1b and Extended Data Fig. 2). Several of these regions provide evidence of positive selection if we assume a model in which the distribution of Neandertal ancestry has been governed by neutral drift; however, this assumption is problematic in light of the evidence for widespread negative selection against Neandertal ancestry reported below (SI 5). As an alternative test for whether Neandertal alleles have been affected by positive selection, we examined the 5% of genes with the highest inferred Neandertal ancestry. We do not detect tissue-specific expression patterns; however genes involved in *keratin filament* formation and some other biological pathways are significantly enriched in Neandertal ancestry in Europeans, East Asians, or both (Extended Data Table 1, SI 6). Thus, Neandertal alleles that affect skin and hair may have been used by modern humans to adapt to non-African environments. We also directly established the relevance of Neandertal alleles to present-day human biology by identifying alleles of Neandertal origin (SI 7), and overlapping this list with alleles that have been associated with phenotype[16]. We identify alleles of Neandertal origin that affect lupus, biliary cirrhosis, Crohn's disease, optic disk size, smoking behavior, IL-18 levels and type 2 diabetes [17] (Extended Data Table 2).

The most striking feature of the introgression map is its large "deserts" of Neandertal ancestry: on a 10 Megabase (Mb) scale on the autosomes, there are 4 windows in Europeans and 14 in East Asians with Neandertal ancestry <0.1% (Extended Data Fig. 2, SI 8). Two analyses show that these deserts are not artifacts of reduced power to detect ancestry. First, when we lower the probability threshold for calling a segment as Neandertal from 90% to 25% our qualitative findings are unchanged (SI 8). Second, when we estimate Neandertal ancestry in regions of low recombination rate where Neandertal haplotypes are longer so that we have more power to detect them, we see a decreased Neandertal ancestry proportion, opposite to the expectation from increased power (ρ_{Spearman}=0.221, P=4.4 × 10^{-4} in Europeans; ρ_{Spearman}=0.226, P=1.9 × 10^{-4} in East Asians) (SI 8). Part of the explanation for the ancestry deserts is likely to be small population sizes shortly after interbreeding, as this could explain why we also observe multi-megabase rises and not just falls in Neandertal ancestry (SI 8). However, selection too appears to have contributed to Neandertal ancestry deserts, as we also detect a correlation to functionally important regions (below).

To explore whether selection provides part of the explanation for regions of reduced Neandertal ancestry, we tested for a correlation of Neandertal ancestry to a previously defined "B-statistic", in which low B implies a high density of functionally important elements[18]. We find that low B is significantly correlated to low Neandertal ancestry: ρ_{Spearman}=0.32 in Europeans (P= 4.9 × 10^{-87}) and ρ_{Spearman}=0.31 in East Asians (P=3.88 × 10^{-68}) (Fig. 2, SI 8). The inference of low Neandertal ancestry in these regions is not an artifact of reduced power, as there is expected to be reduced genetic variation in regions of low B which should make introgressed Neandertal haplotypes stand out more clearly (Extended Data Table 3). We also estimated Neandertal ancestry in quintiles of B-statistic using an approach that is not biased by varying mutation rates, recombination rates, or genealogical tree depth[19], and confirmed that the quintile with the highest B has significantly higher Neandertal ancestry than the other quintiles (P=7×10^{-4}) (Extended Data Table 4, SI 9).

The largest deserts of Neandertal ancestry are on chromosome X, where the mean Neandertal ancestry is about a fifth of the autosomes (Table 1). The power of our CRF to detect Neandertal ancestry is higher on chromosome X than on the autosomes (Extended Data Table 3), implying that this observation cannot be an artifact of reduced power. At least some of the reduction in Neandertal ancestry that we observe on chromosome X must be due to selection, since just as on the autosomes, we observe that Neandertal ancestry is positively correlated with B-statistic (ρ_{Spearman}=0.276, P = 3.1 × 10^{-4} for Europeans; ρ_{Spearman}=0.176, P = 0.02 for East Asians) (Fig. 2, SI 8). Studies in many species have shown that genes responsible for reduced male fertility disproportionally map to chromosome X[20–22]. We hypothesized that this "Large X Effect" [23] could explain why chromosome X was more resistant to introgression of Neandertal ancestry than the autosomes.

If male hybrid sterility is contributing to our observations, a prediction is that the responsible genes will be disproportionally expressed in testis[24]. To test this hypothesis, we analyzed gene transcripts from 16 human tissues[25] and defined "tissue-specific" genes as those with a significantly higher expression level in that tissue than any other. We found that only genes specifically expressed in testis were enriched in regions of low Neandertal ancestry, an effect that remained significant after permuting gene annotations while preserving the correlation structure between Neandertal ancestry and gene expression (P = 0.0095 in Europeans; P = 0.018 in East Asians) (Table 2, SI 6). However, hybrid sterility is not the only factor responsible for selection against Neandertal material, as Neandertal ancestry is also depleted in conserved pathways such as RNA processing ($P<0.05$; Extended Data Table 2; SI 6).

We have shown that interbreeding of Neandertals and modern humans introduced alleles onto the modern human genetic background that were not tolerated and were swept away, in part because they contributed to male hybrid sterility. The resulting reduction in Neandertal ancestry was quantitatively large: in the fifth of the genome with highest B, Neandertal ancestry is 1.54 ± 0.15 times the genome-wide average (Extended Data Table 4; SI 9)[19]. If we assume that this subset of the genome was unaffected by selection, this implies that the proportion of Neandertal ancestry shortly after introgression must have been >3% rather than the present-day ~2%. In passing, we note that the large effect of negative selection on present-day levels of Neandertal ancestry may explain why the proportion of Neandertal ancestry is significantly higher in present-day East Asians than in Europeans (Table 1) [7,14]; population sizes appear to have been smaller in East Asians than Europeans for some of the time since their separation[26], and this would result in less efficient selection to remove Neandertal-derived deleterious alleles. The evidence for male hybrid sterility is particularly remarkable when compared with mixed populations of present-day humans in which no convincing signals of selection against alleles inherited from one of the mixing populations have been found despite high power to detect such effects[27]. Thus, while the time of separation between Neandertals and modern humans was about 5 times larger than that between present-day Europeans and West Africans[2], the biological incompatibility was far greater. A potential explanation is the "snowball effect", whereby hybrid sterility genes are expected to accumulate in proportion to the square of the substitutions between two taxa because two interacting loci need to change to produce an incompatibility ("Dobzhansky-Muller incompatibilities")[28]. An important direction for future work is to explore whether similar phenomena have affected other interbreeding events between diverged humans.

Online Methods

Conditional Random Field for inferring Neandertal local ancestry

Consider a haploid genome in a test population that carries Neandertal ancestry, for example Europeans. Given the allelic states of a sequence of SNPs along this haplotype, we would like to infer the ancestral state of the allele at each SNP, specifically, whether it has entered modern humans through Neandertal gene flow. In addition to the test haplotype, the data we analyze consist of a panel of haplotypes from the sub-Saharan African Yoruba (YRI) who we assume harbor no Neandertal ancestry [1]. To determine the allelic state of the

Neandertals, we use a high-coverage Neandertal genome [2]. We determine the ancestral and derived allele at each SNP using a 6-primate consensus sequence [13]. To estimate the genetic distance between adjacent SNPs, we use the Oxford combined linkage disequilibrium (LD) map [29]. We specify the distribution of the unobserved Neandertal ancestry states at each SNP given the observed genetic data as a Conditional Random Field (CRF) [9]. Intuitively, we specify feature functions that relate the observed data and the unobserved ancestral state at each SNP ("emission functions") as well as feature functions that relate the unobserved ancestral states at adjacent SNPs ("transition functions"). Thus, the model is a linear-chain CRF. The feature functions and their associated parameters fully specify the distribution of the unobserved ancestral states given the observed data. Given the parameters and the observed data, we are able to infer the marginal probability of Neandertal ancestry at each SNP of the haploid genome. We compute the marginal probabilities efficiently using the forward-backward algorithm [9,30]. SI 1 presents the mathematical details.

Feature functions

The emission functions couple the unobserved ancestral state at a SNP to the observed features. We use two classes of emission functions.

The first class of emission function captures information from the joint patterns observed at a single SNP across Europeans, Africans and Neandertals. These features are indicator functions that assume the value "1" when a specific pattern is observed at a SNP and "0" otherwise. We use feature functions that pick out two classes of allelic patterns. One of these features is 1 if at a given SNP, the test haplotype carries the derived allele, all the YRI haplotypes carry the ancestral allele, and either of the two Neandertal alleles is derived. SNPs with this joint configuration have an increased likelihood of Neandertal ancestry. In the CRF, an increased likelihood associated with this feature is reflected in the fact that the parameter is positive with a magnitude determined by the informativeness of the feature. The second feature is 1 if at a given SNP, the test haplotype carries a derived allele that is polymorphic in the panel of Africans but absent in the Neandertal. SNPs with this joint configuration have a decreased likelihood of Neandertal ancestry.

The second class of emission functions uses multiple SNPs to capture the signal of Neandertal ancestry. Specifically, we compare the divergence of the test haplotype to the Neandertal sequence to the minimum divergence of the test haplotype to all African haplotypes over non-overlapping 100 kilobase (kb) windows (the size scale we expect for Neandertal haplotypes today based on the time of Neandertal gene flow into modern humans [10]). In a region of the genome where the test haplotype carries Neandertal ancestry, we expect the test haplotype to be closer to the Neandertal sequence than to most modern human sequences (albeit with a large variance), and we expect the pattern to be reversed outside these regions. While computing distance to the Neandertal sequence, we build a Neandertal haplotype by choosing the derived allele at heterozygous sites so that this distance is effectively the minimum distance of the potentially introgressed test haplotype to one of the two Neandertal haplotypes.

The transition feature function modulates the correlation of the ancestral states at adjacent SNPs. We define this feature function as an approximation, at small genetic distance, to the log of the transition probabilities of a standard Markov process of admixture between two populations. This approximation makes parameter estimation in the CRF efficient.

Parameter Estimation

To estimate the parameters of the CRF, we need haplotypes labeled with Neandertal ancestries; that is, training data. Since we do not in fact know the true Neandertal state in any individual, we estimated the CRF parameters on data simulated under a demographic model. We estimate parameters by maximizing the L2-regularized conditional log likelihood using a limited-memory version of LBFGS [31]. We fixed the value of the parameter associated with the L2-penalty at 10 although a broad range of values appear to work in practice. We assumed a simple demographic model relating Africans, Europeans and Neandertals with Neandertal-modern human admixture occurring 1,900 generations ago [10] and a fraction of Neandertal ancestry of 3% [1]. The model parameters were broadly constrained by the observed allele frequency differentiation between the West African YRI and European American CEU populations and by the observed excess sharing of alleles between Europeans and Africans relative to Neandertals. The simulations incorporated hotspots of recombination [11] as well as the reduced power to detect low-frequency alleles from low-coverage sequencing data [12].

Validation of the CRF

We assessed the accuracy of the CRF to predict Neandertal ancestry using simulated data. Given the marginal probabilities estimated by the CRF, we estimated the precision (fraction of predictions that are truly Neandertal) and the recall (fraction of true Neandertal alleles that are predicted) as we vary the threshold on the marginal probability for an allele to be declared Neandertal. We also evaluated the accuracy when the haplotype phase needs to be inferred and when the genetic map has errors. Since the CRF parameter estimation assumes a specific demographic model, we were concerned about the possibility that the inferences might be sensitive to the model assumed. We therefore perturbed each demographic parameter in turn and applied the CRF to data simulated under these perturbed models, fixing the parameters of the CRF to the estimates obtained under the original model. For each of these perturbed models, we evaluated the false discovery rate (1-precision) when we restrict to sites at which the CRF assigns a marginal probability of at least 0.90. SI 2 presents the details.

Preparation of 1000 Genomes Data

We applied the CRF to the computationally phased haplotypes in each of the 13 populations in the 1000 Genomes project [12] (1KG), excluding the west African Yoruba (YRI). The CRF requires reference genomes from Africans and Neandertals. For the African population, we used 176 haplotypes from 88 YRI individuals. For the Neandertal genome, we used the genotypes called from the recently generated high-coverage Neandertal sequence [2]. We restricted our analysis to sites passing the filters described in ref. [2] and for which the genotype quality score GQ ≥ 30. These filters discard sites that are identified as repeats by the Tandem Repeat Finder [32], that have Phred-scaled mapping quality scores of MQ < 30, or that map to regions where the alignment is ambiguous or which fall within the upper or lower 2.5th percentile of the sample-specific coverage distribution (applied within the regions of unique mappability binned according to the GC-content of the reference genome). For the mappability filter, we used the liberal filter that requires that at least 50% of all 35-mers that overlap a position do not map to any other position in the genome allowing up to one mismatch. We further restricted our analysis to sites that are biallelic across the Neandertal and the 1000 Genomes project samples. For each haplotype analyzed, we also restricted to the set of polymorphic sites in the population containing the haplotype. After filtering, we were able to analyze 26,493,206 SNPs on the autosomes and 817,447 SNPs on chromosome X. We obtained genetic distances from the Oxford combined LD map lifted over to hg19 coordinates [29]. For the X chromosome, we obtained an appropriate sex-averaged map by scaling the X chromosome LD-based map by 2/3.

Statistics for measuring Neandertal ancestry

We computed several statistics to summarize the Neandertal ancestry inferred by the CRF. We estimated the proportion of an individual diploid genome that is confidently inferred to be Neandertal as the fraction of sites for which the marginal probability is ≥ 90%. To assess variation in the proportion of Neandertal ancestry along the genome, we computed the fraction of alleles across individuals with marginal probability greater than a specified threshold. This statistic is likely to be affected by variation in power along the genome. Hence, we also consider an estimate of the ancestry proportion obtained by averaging the marginal probability across individuals. Depending on the analyses, these statistics are estimated at a single SNP or in non-overlapping windows of a specified size.

To assess whether the predictions made by the CRF are sensible, we inferred Neandertal ancestry using the low-coverage genome from the Vindija Neandertals [1]. For this analysis, we restricted to sites at which there is at least one read with mapping quality score between 60 and 90 and base quality of at least 40. As a second validation analysis, we applied the CRF to the sub-Saharan African Luhya (LWK), using the parameters optimized for non-Africans. We empirically assessed the accuracy of the CRF on the 1000 Genomes project data by assuming that LWK has no Neandertal ancestry, that the false discovery rate in each non-African population is equal to the false discovery rate in LWK, and using the genome-wide proportion of Neandertal ancestry estimated in ref. [2] (SI 3). We computed the theoretical standard deviation in the proportion of Neandertal ancestry [33] assuming a pulse model of admixture with 2% Neandertal ancestry followed by 2,000 generations of random mating, and 2.03 Gigabases as the number of bases of the high-coverage Neandertal genome that pass filters.

Tiling path of Neandertal haplotypes

We identified Neandertal haplotypes as runs of consecutive alleles along a test haplotype assigned a marginal probability of > 90%. We filtered haplotypes smaller than 0.02 cM. At each SNP that is covered by at least one such haplotype, we estimated the allelic state as the consensus allele across the spanning haplotypes. See SI 4.

Functional analysis of introgressed alleles

We defined two subsets of Consensus Coding Sequence (CCDS) genes [34]. We define a gene with “low Neandertal ancestry” as one in which all alleles across all individuals have a marginal probability ≤ 10%. We also require that the genes included in this analysis include at least ≥ 100 SNPs within a 100 kb window centered at its midpoint (this excluded genes with low power). We define a gene with "high Neandertal ancestry" as one that is in the top 5% of CCDS genes ranked by the average marginal probability across individual haplotypes.

Functional enrichment analysis

We tested for enrichment of Gene Ontology (GO) [35] categories in genes with low or high Neandertal ancestry, using the hypergeometric test implemented in the FUNC package [36]. We report multiple-testing corrected P-values estimated from 1000 permutations for the GO enrichment analysis (Family-Wise Error Rate – FWER). Given the observed correlation between Neandertal ancestry and B-statistic [18], a concern is that the functional categories may not be randomly distributed with respect to B-statistic. To control for this, we assigned a B-statistic to each gene (estimated as an average of the B-statistic over the length of the gene) as well as a uniform random number. This resulted in 17,249 autosomal genes. Genes were binned into 20-equal sized bins based on the gene-specific B-statistic. Within each bin, genes were sorted by Neandertal ancestry and then by the random number. Genes ranked within the top 5% within each bin were used for the analysis. See SI 6 for details.

Identifying alleles born in Neandertals and cross-correlation with association study data

To infer whether an allele segregating in a present-day human population was introduced by Neandertal gene flow, we defined a probable Neandertal allele as one with marginal probability of ≥ 90% and a non-Neandertal allele as having a marginal probability of ≤ 10%. A SNP at which all of the confident non-Neandertal alleles as well as all alleles in YRI are ancestral and all of the confident Neandertal alleles are derived is inferred to be of Neandertal origin. This allows for some false negatives in the prediction of the CRF. This procedure yields 97,365 Neandertal-derived SNPs when applied to the predictions in Europeans and East Asians. We downloaded the variants listed in the NHGRI GWAS catalog [16], retaining entries for which the reported association is a SNP with an assigned rs-number, and where the nominal P-value is $<5\times10^{-8}$. This resulted in 5,022 associations, which we then intersected with the Neandertal-derived list. See SI 7 for details.

Identification of genomic regions deficient in Neandertal ancestry

We measured the fraction of alleles across individuals and SNPs that have less than a specified proportion of Neandertal ancestry measured within non-overlapping 10 Megabase (Mb) windows. We chose a threshold of marginal probability of >25% as this threshold was found to lead to high recall in our empirical assessment (SI 3). We reported windows for which this statistic is < 0.1%.

To understand the causes for variation in Neandertal ancestry, we tested for correlation to a B-statistic. Each SNP was annotated with the B-statistic lifted over to hg19 coordinates. We assessed correlation between B and estimates of Neandertal ancestry proportion at a nucleotide level as well as at different size scales, and separately on the autosomes and chromosome X (SI 8). We also used wavelet decomposition[37] to analyze the correlation of the inferred Neandertal ancestry between Europeans and East Asians at multiple size scales (SI 10). Figure 2 reports the relation between the mean marginal probability of Neandertal ancestry across individuals and quintiles of B-statistic at each SNP. To assess significance, we estimated Spearman's correlation ρ_{Spearman} and standard errors using a block jackknife [38] with 10 Mb blocks (SI 8).

To understand the contribution of demography to variation in Neandertal ancestry along the genome, we measured the coefficient of variation, at a 10 Mb size scale, of the proportion of ancestry estimated as defined above. We then applied the CRF to data simulated under diverse demographic models and compared the coefficient of variation to the observed value (SI 8).

Power to infer Neandertal ancestry as a function of demography and genomic features

To assess the power of the CRF to infer Neandertal ancestry, we simulated data under diverse demographic models [39]. In one simulation series, we varied effective population size to approximate the effect of background selection and measured recall at a precision of 90%. In a second series, we assessed power on chromosome X versus the autosomes by matching the effective population size, recombination rate and mutation rate to estimated values for chromosome X. See SI 2.

Unbiased estimate of the proportion of Neandertal ancestry as a function of B-statistic

To estimate the proportion of Neandertal ancestry in an unbiased way, we divided the genome into quintiles of B, and estimated the proportion of Neandertal ancestry using a statistic first published in ref. [19]. This statistic measures how much closer a non-African individual is to Denisova than an African individual, divided by the same quantity replacing the non-African individual with Neandertal. We report the estimated proportion of Neandertal ancestry in each quintile as a fraction of the genome-wide mean and obtain standard errors using a block jackknife with 100 blocks.

We analyzed data from 27 deeply sequenced genomes: 25 present-day humans and the high-coverage Neandertal and Denisova [14] genomes. For each, we required that sites pass the more stringent set of the two filters described in ref. [2], have a genotype quality of $GQ \geq 45$, and have an ancestral allele that can be determined based on comparison to chimpanzee and at least one of gorilla or orangutan. We computed a Z-score for the difference in the ancestry across the bin of highest B-statistic versus the rest and used a Bonferroni correction for ten hypotheses (5 hypotheses based on which set of bins we merge and a 2-sided test in each). In our main analysis, we analyzed both transitions and transversions and pooled genomes for all non-African samples. We also analyzed other subsets of the data: transversions only in non-Africans, in Europeans and in eastern non-Africans. See SI 9 for details.

Tissue-specific expression

We defined tissue-specific expression levels using the Illumina BodyMap 2.0 RNA-seq data [25], which contains expression data from 16 human tissues. We identified genes that are expressed in a tissue-specific manner using the DESeq package [40] and used a P-value cutoff of 0.05. We tested enrichment of tissue-specific genes in regions of high or low Neandertal ancestry. A concern when testing for enrichment is that clustering of similarly expressed genes coupled with the large size of regions of low Neandertal ancestry might lead to spurious signals of enrichment. Hence, we devised a permutation test that randomly rotated the annotations of genes (treating each chromosome as a circle) while maintaining the correlation within genes and within Neandertal ancestry as well as between Neandertal ancestry and genes. We tested enrichment on the whole genome, on the autosomes alone, and on chromosome X alone. We generated 1,000 random rotations for each test except for the X chromosome for which we generated all possible rotations. We computed the fraction of permutations for which the P-value of Fischer's exact test is at least as low as the observed P-value (See SI 6).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Adrian Briggs, Priya Moorjani, Molly Przeworski, Daven Presgraves, Amy Williams, and two anonymous reviewers for critical comments, and Kathryn Kavanagh for help with Extended Figure 2. We are grateful for support from the Presidential Innovation Fund of the Max Planck Society, NSF HOMINID grant #1032255 and NIH grant GM100233. SS was supported by a post-doctoral fellowship from the Initiative for the Science of the Human Past at Harvard. DR is a Howard Hughes Medical Institute Investigator.

References

1. Green RE, et al. A draft sequence of the Neandertal genome. Science. 2010; 328:710–722. [PubMed: 20448178]
2. Prüfer K, et al. The complete genome sequence of a Neandertal from the Altai Mountains. Nature. 2014; 505:43–49. [PubMed: 24352235]
3. Abi-Rached L, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science. 2011; 334:89–94. [PubMed: 21868630]
4. Mendez FL, Watkins JC, Hammer MF. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am J Hum Genet. 2012; 91:265–274. [PubMed: 22883142]
5. Mendez FL, Watkins JC, Hammer MF. Neandertal origin of genetic variation at the cluster of OAS immunity genes. Mol Biol Evol. 2013; 30:798–801. [PubMed: 23315957]
6. Yotova V, et al. An X-linked haplotype of Neandertal origin is present among all non-African populations. Mol Biol Evol. 2011; 28:1957–1962. [PubMed: 21266489]
7. Wall JD, et al. Higher levels of neanderthal ancestry in East aSians than in Europeans. Genetics. 2013; 194:199–209. [PubMed: 23410836]
8. Lachance J, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell. 2012; 150:457–469. [PubMed: 22840920]
9. Lafferty, J.; McCallum, A.; Pereira, FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning; 2001. p. 282-289.
10. Sankararaman S, Patterson N, Li H, Paabo S, Reich D. The date of interbreeding between Neandertals and modern humans. PLoS Genet. 2012; 8:e1002947. [PubMed: 23055938]
11. Hellenthal G, Stephens M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics. 2007; 23:520–521. [PubMed: 17150995]
12. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
13. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008; 18:1814–1828. [PubMed: 18849524]

14. Meyer M, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012; 338:222–226. [PubMed: 22936568]

15. 23andMe White Paper 23-05. Neanderthal Ancestry Estimator. <https://23andme.https.internapcdn.net/res/pdf/hXitekfSJe1lcIy7-Q72XA_23-05_Neanderthal_Ancestry.pdf>

16. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]

17. The SIGMA Type 2 Diabetes Consortium. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature. 2014; 506(7486):97–101. [PubMed: 24390345]

18. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009; 5:e1000471. [PubMed: 19424416]

19. Reich D, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010; 468:1053–1060. [PubMed: 21179161]

20. Tucker PK, Sage RD, Wilson AC, Eichler EM. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution Int J Org Evolution. 1992; 46:1146–1163.

21. Good JM, Dean MD, Nachman MW. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. Genetics. 2008; 179:2213–2228. [PubMed: 18689897]

22. Presgraves DC. Sex chromosomes and speciation in Drosophila. Trends Genet. 2008; 24:336–343. [PubMed: 18514967]

23. Coyne, JAO.; HA. Speciation and its Consequences. Otte, D.; Endler, JA., editors. 1989. p. 180-207.

24. Wu CI, Davis AW. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic basis. American Naturalist. 1993; 142:187–212.

25. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]

26. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet. 2007; 39:1251–1255. [PubMed: 17828266]

27. Bhatia, G., et al. Genome-wide scan of 29,141 African Americans finds no evidence of selection since admixture. arXiv:1312.2675 [q-bio.PE]

28. Orr HA, Turelli M. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution. 2001; 55:1085–1094. [PubMed: 11475044]

29. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005; 310:321–324. [PubMed: 16224025]

30. Sutton, C.; McCallum, A. Introduction to Statistical Relational Learning. Getoor, Lise; Taskar, Ben, editors. Vol. Ch. 4. MIT Press; 2007. p. 93-128.

31. Byrd RH, Nocedal J, Schnabel RB. Representations of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming. 1994; 63.1-3:129–156.

32. Tandem Repeat Finder. <url={http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz}>

33. Gravel S. Population genetics models of local ancestry. Genetics. 2012; 191:607–619. [PubMed: 22491189]

34. Pruitt KD, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009; 19:1316–1323. [PubMed: 19498102]

35. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

36. Prufer K, et al. FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics. 2007; 8:41. [PubMed: 17284313]

37. Percival, DB.; Walden, AT. Wavelet Methods for Time Series Analysis. Cambridge University Press; 2005.

38. Kunsch HR. The jackknife and the bootstrap for general stationary observations. The Annals of Statistics. 1989; 17:1217–1241.

39. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–338. [PubMed: 11847089]

40. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]

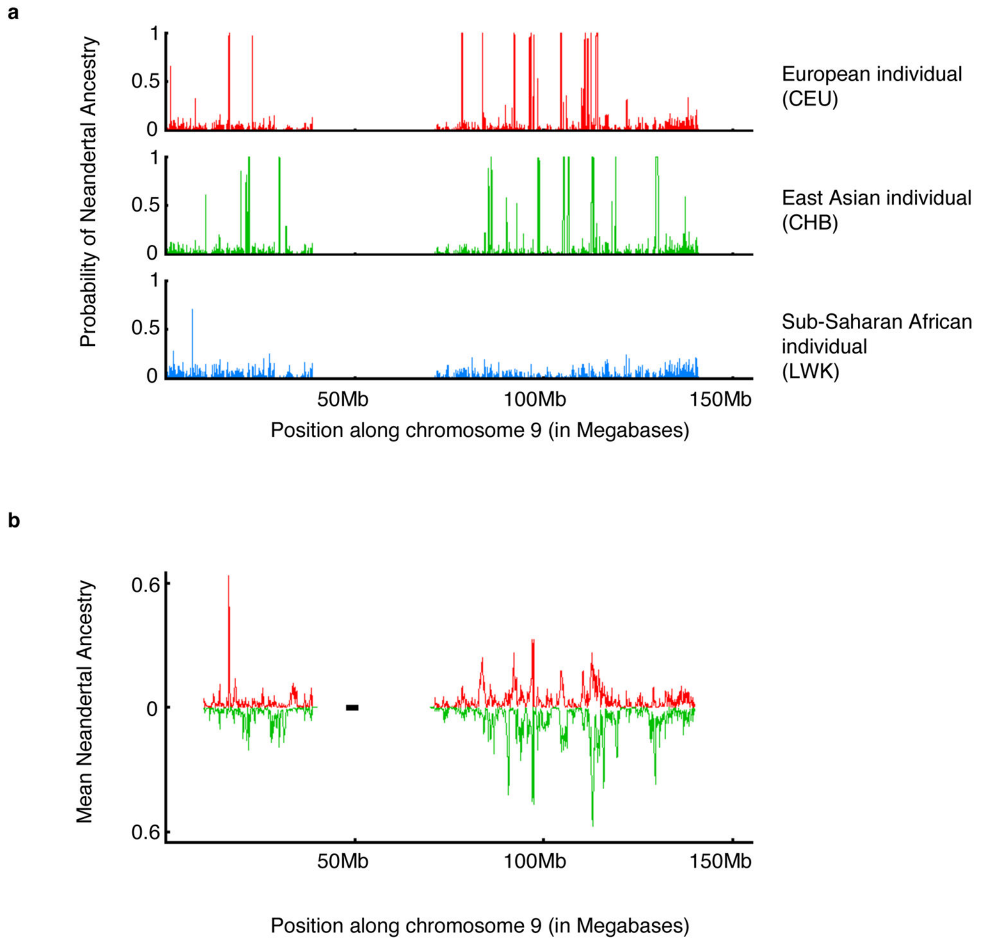


Figure 1. Maps of Neandertal ancestry

(**a**) Individual maps: We show the marginal probability of Neandertal ancestry for 1 European American, 1 East Asian and 1 sub-Saharan African phased genome across chromosome 9. (**b**) Population maps: We show the average of the inferred proportion of Neandertal ancestry in Europeans (above axis) and East Asians (below axis) in non-overlapping 100 kb windows on chromosome 9.

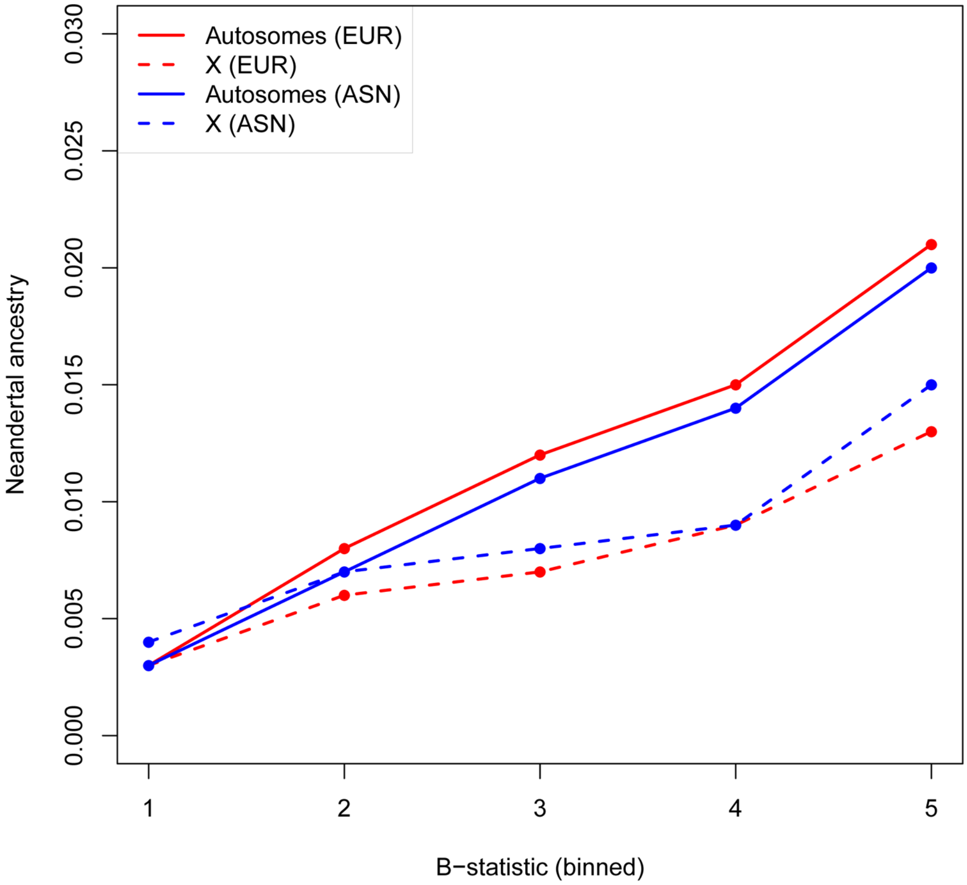
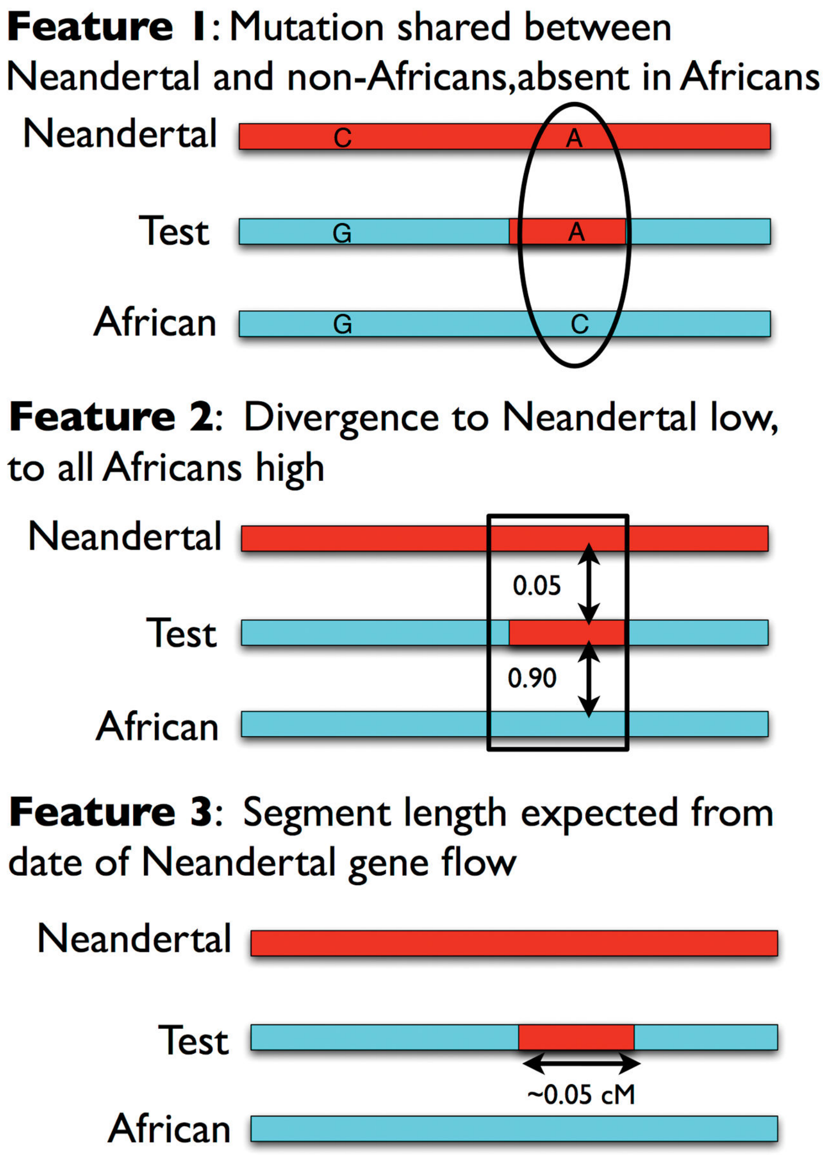


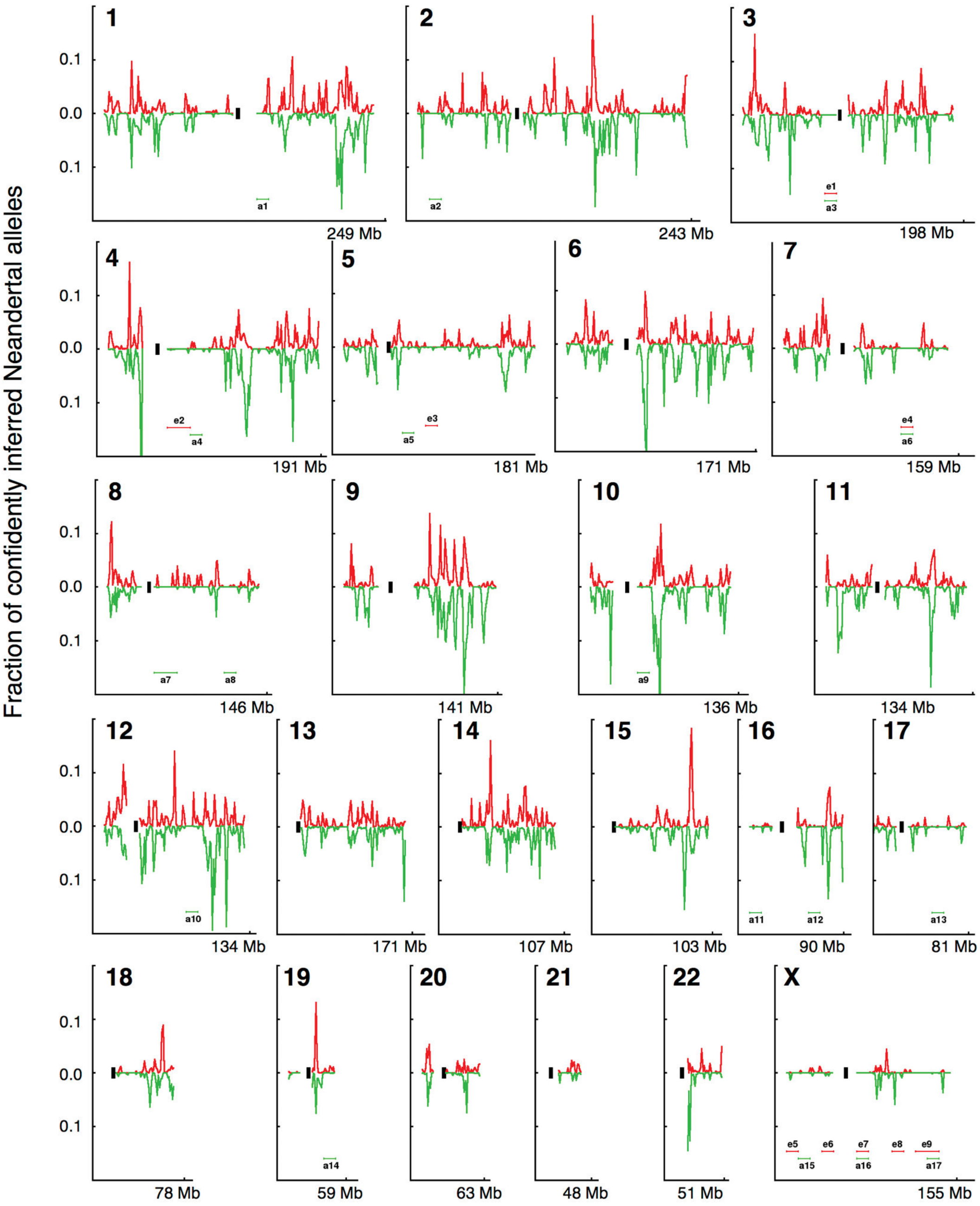
Figure 2. Functionally important regions are deficient in Neandertal ancestry

We plot the median of the proportion of Neandertal ancestry (the average over the marginal probability of Neandertal ancestry assigned to each individual allele at a SNP) within quintiles of a B-statistic that measures proximity to functionally important regions (1-low, 5-high). We show results on the autosomes and chromosome X, and in Europeans and East Asians.


Feature 1: Mutation shared between Neandertal and non-Africans,absent in Africans
Neandertal
C
A
Test
G
A
African
G
C
Feature 2: Divergence to Neandertal low, to all Africans high
Neandertal
0.05
Test
0.90
African
Feature 3: Segment length expected from date of Neandertal gene flow
Neandertal
Test
~0.05 cM
African

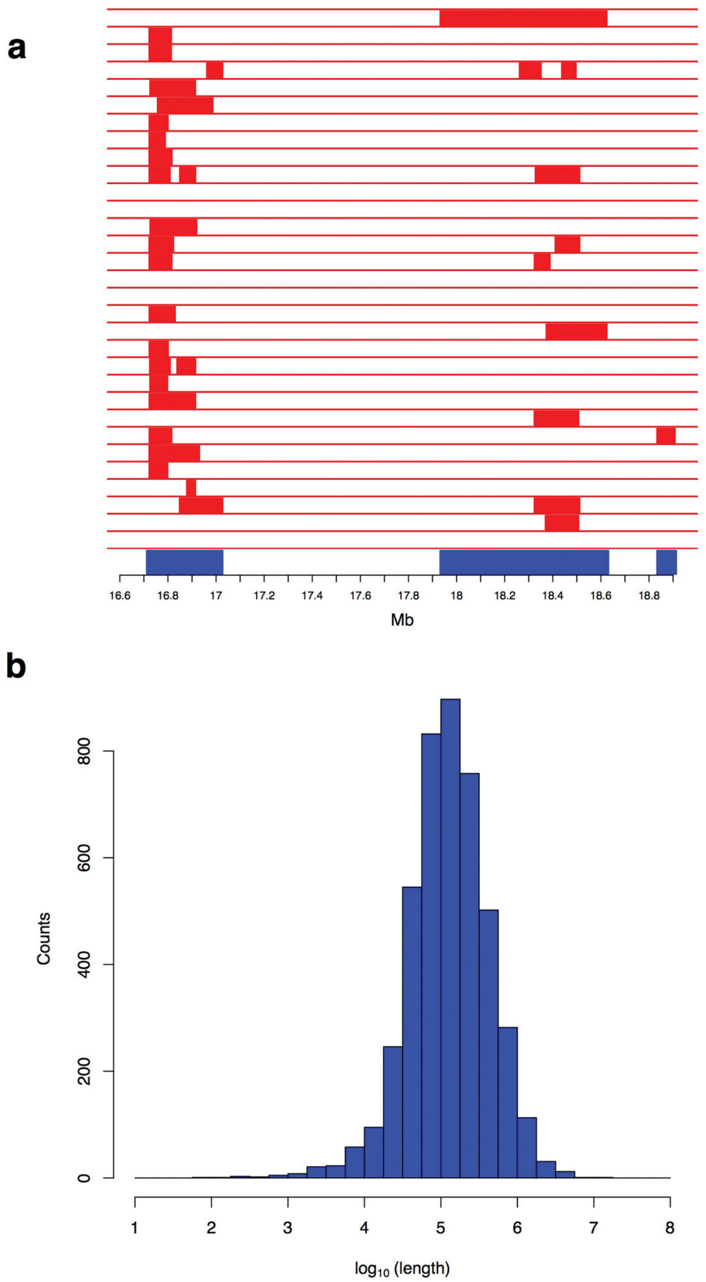
Extended Data Figure 1. Three features used in the Conditional Random Field for predicting Neandertal ancestry

Feature 1: Patterns of variation at a single SNP. Sites where a panel of sub-Saharan Africans carries the ancestral allele and where the sequenced Neandertal and the test haplotype carry the derived allele are likely to be derived from Neandertal gene flow. Feature 2: Haplotype divergence patterns. Genomic segments where the divergence of the test haplotype to the sequenced Neandertal is low while the divergence to a panel of sub-Saharan Africans is high are likely to be introgressed. Feature 3: we search for segments that have a length consistent with what is expected from the Neandertal-into-modern human gene flow 2,000 generations ago, corresponding to a size of about 0.05cM = (100cM/Morgan)/(2000 generations).



Extended Data Figure 2. Map of Neandertal ancestry in 1000 Genomes European and East Asian populations

For each chromosome, we plot the fraction of alleles confidently inferred to be of Neandertal origin (probability > 90%) in non-overlapping 1 Mb windows. We label 10 Mb scale windows that are deficient in Neandertal ancestry (e1–e9, a1–a17) (SI 8).



Extended Data Figure 3. Tiling path from confidently inferred Neandertal haplotypes

a, Example tiling path at the *BNC2* locus on chromosome 9 in Europeans. Red denotes confident Neandertal haplotypes. Blue denotes the resulting tiling path. We identified Neandertal haplotypes by scanning for runs of consecutive SNPs along a haplotype with a marginal probability > 90% and requiring the haplotypes to be at least 0.02 cM long. **b**, Distribution of contig lengths obtained by constructing a tiling path across confidently inferred Neandertal haplotypes. On merging Neandertal haplotypes in each of the 1000 Genomes European and East Asian populations, we reconstructed 4,437 Neandertal contigs with median length 129 kb.

Table 1

Genome-wide estimates of Neandertal ancestry

			Neandertal ancestry (%)	
	Population	**Individuals**	**Autosomes**	**X**
Europeans	CEU	85	1.17±0.08	0.21±0.17
	FIN	93	1.20±0.07	0.19±0.14
	GBR	89	1.15±0.08	0.20±0.15
	IBS	14	1.07±0.06	0.23±0.18
	TSI	98	1.11±0.07	0.25±0.20
East Asians	CHB	97	1.40±0.08	0.30±0.21
	CHS	100	1.37±0.08	0.27±0.21
	JPT	89	1.38±0.10	0.26±0.21
Americans	CLM	60	1.14±0.12	0.22±0.16
	MXL	66	1.22±0.09	0.21±0.15
	PUR	55	1.05±0.12	0.20±0.15
Africans	LWK	97	0.08±0.02	0.04±0.07
	ASW	61	0.34±0.22	0.07±0.11

Note: For each computationally phased genome in each population, we estimated the probability of Neandertal ancestry at each SNP and the fraction of autosomal and X-chromosome SNPs that are confidently Neandertal in each individual (marginal probability > 90%). The table reports the average and standard deviation of this measure across individuals within each population.

Table 2

Enrichment of tissue-specific genes in regions deficient in Neandertal ancestry

	Europeans			East Asians		
Tissue	**Genome**	**Chr. X**	**Autosomes**	**Genome**	**Chr. X**	**Autosomes**
adipose	0.93	1	0.81	0.99	1	0.95
adrenal	0.5	NA	0.5	0.42	NA	0.42
blood	0.99	0.98	0.99	0.94	0.73	0.94
brain	1	1	1	1	1	1
breast	0.98	0.63	0.99	1	0.94	1
colon	0.64	0.77	0.63	0.94	0.97	0.89
heart	0.99	0.71	0.99	0.8	0.57	0.81
kidney	1	0.15	1	1	0.08	1
liver	0.99	0.99	0.99	1	0.86	1
lung	0.96	0.64	0.96	0.99	0.87	0.99
lymph	0.88	0.62	0.9	0.99	0.51	0.99
ovary	0.84	0.95	0.81	0.62	0.91	0.58
prostate	1	0.79	1	1	0.73	1
muscle	0.95	0.7	0.95	0.83	0.1	0.88
testes	0.0095	0.13	0.016	0.018	0.039	0.055
thyroid	0.86	0.62	0.88	0.87	0.94	0.86

Note: We compare tissue-specific genes (defined as those that are significantly more highly expressed in the specified tissue than in any of the 15 other tissues) to all other expressed genes in that tissue. Of the sixteen tissues tested, only testis-specific genes are significantly enriched in the regions deficient in Neandertal ancestry, defined as locations where all sites across all individuals are assigned a marginal probability of Neandertal ancestry of <10% (47% of genes in Europeans and 52% of genes in East Asians fall in this category). NA denotes that there were no tissue-specific genes for this tissue on chromosome X.

Extended Data Table 1

Gene categories enriched or depleted in Neandertal ancestry. Enrichment of Gene Ontology categories in genes with depleted or elevated Neandertal ancestry was assessed using the hypergeometric test implemented in the FUNC package. We report Family-wise error rate P-values (FWER) associated with each GO category (P-values corrected for the testing of multiple categories).

Biological pathway (GO categorization)	Neandertal ancestry	Europe FWER	East Asian FWER
nucleic acid binding (molecular_function, GO:0003676)	Depleted	0.018	0.032
RNA processing (biological_process, GO:0006396)	Depleted	0.004	0.049
ribonucleoprotein complex (cellular_component, GO:0030529)	Depleted	<0.001	0.027
organelle part (cellular_component, GO:0044422)	Depleted	<0.001	0.037
intracellular organelle part (cellular_component, GO:0044446)	Depleted	<0.001	0.025
mRNA metabolic process (biological_process, GO:0016071)	Depleted	<0.001	0.014
nuclear lumen (cellular_component, GO:0031981)	Depleted	0.039	0.017
nuclear part (cellular_component, GO:0044428)	Depleted	0.005	0.022
keratin filament (cellular_component, GO:0045095)	Enriched	<0.001	<0.001

Extended Data Table 2

Neandertal derived alleles that have been reported to be associated with phenotypes in genome-wide association studies (GWAS). We identified alleles that are likely to have been introduced by Neandertal gene flow (SI 10) and intersected these alleles with SNPs that have been shown to be associated with phenotypes from the NHGRI GWAS catalog as well as from a recent GWAS for type 2 diabetes [17]).

rs id	Coordinates	Derived allele	Derived allele frequency (%)		Phenotype
			Europeans	East Asians	
rs12531711	7:128,617,466	G	10.03	0.17	Systemic lupus erythematosus, Primary biliary cirrhosis
rs3025343	9:136,478,355	A	8.44	0.00	Smoking behavior
rs7076156	10:64,415,184	A	26.52	8.74	Crohn’s disease
rs12571093	10:70,019,371	A	16.35	14.86	Optic disc size
rs1834481	11:112,023,827	G	21.50	0.35	Interleukin-18 levels
rs11175593	12:40,601,940	T	1.98	3.32	Crohn’s disease
rs75493593	17:6,945,087	T	1.85	12.06	
rs75418188	17:6,945,483	T	1.85	11.54	Type-2 Diabetes
rs117767867	17:6,946,330	T	1.85	11.54	

Extended Data Table 3

Power to infer Neandertal ancestry. a, Simulated power to infer Neandertal ancestry as a function of the effective population size. **b**, Power to infer Neandertal ancestry on chromosome X versus the autosomes. Recall is computed at a precision of 90%. Standard errors are estimated by a block jackknife with 100 blocks.

a

Effective population size	**Recall**
2500	0.552 ± 0.009
5000	0.506 ± 0.009
7500	0.430 ± 0.006
10000	0.384 ± 0.006

b

	Recall
Autosomes	0.384 ± 0.006
X	0.495 ± 0.009

Extended Data Table 4

Unbiased estimate of the proportion of Neandertal ancestry as a function of B-statistic. We estimate the proportion of Neandertal ancestry in quintiles of B-statistics. We find a significant excess of Neandertal ancestry in the quintile with the highest B-statistic relative to the remaining four quintiles (significant after correcting for ten hypotheses).

	11 Non-Africans (transitions + transversions)		**11 Non-Africans (transversions only)**		**4 Europeans (transversions only)**		**7 Eastern (transversions only)**	
	Est.	**Err.**	**Est.**	**Err.**	**Est.**	**Err.**	**Est.**	**Err.**
Quintile 1: B=0–0.63	0.641	0.304	0.672	0.316	0.472	0.397	0.778	0.317
Quintile 2: B=0.63–0.80	0.825	0.209	0.779	0.234	0.849	0.290	0.750	0.236
Quintile 3: B=0.80–0.88	0.578	0.248	0.745	0.298	0.987	0.349	0.647	0.297
Quintile 4: B=0.88–0.94	0.684	0.184	0.676	0.208	0.446	0.256	0.771	0.221
Quintile 5: B=0.94–1.00	1.537	0.152	1.445	0.164	1.502	0.185	1.419	0.177
B≥0.94 vs. B ≤ 0.94	Z=3.82		Z=3.02		Z=3.12		Z=2.58	
Correct for 10 hypotheses	P=0.00066		P=0.013		P=0.0090		P=0.049	