Map of Neandertal ancestry : Supporting Information

December 12, 2013

List of Supplementary Figures

List of Supplementary Tables

Contents

Analysis of the genomes of archaic hominins, such as Neandertals and Denisovans, has revealed that these groups have contributed to the genetic variation of modern human populations. Yet, we know little about how these ancient mixtures have shaped the genetic structure of human populations and about their impact on human evolution although studies have begun to investigate the phenotypic impact of these mixtures at specific loci (Abi-Rached et al., 2011; Mendez et al., 2012, 2013; Yotova et al., 2011). To answer these questions systematically, we need a map of archaic ancestry i.e., a map that labels whether each region of an individual genome is descended from an archaic population. Building such a map is technically challenging because of the antiquity of these gene flow events.

We describe a computational method that can be used to infer regions of archaic ancestry using patterns of genetic variation in modern humans and archaic hominins. We apply this method to data from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) and a recently sequenced high-coverage Neandertal (Prüfer et al., 2013) to build a map of Neandertal ancestry in modern non-Africans. The method is more generally applicable to inferring regions derived from other ancient admixtures such as the admixture with Denisovans that contributed genetic material to the Melanesian populations (Reich et al., 2010; Meyer et al., 2012).

Inferring local ancestry in recently admixed populations is a well-studied problem. A number of computational methods have been developed for this task and have been shown to accurately infer local ancestries (Price et al., 2009; Tang et al., 2006; Sankararaman et al., 2008; Johnson et al., 2011; Brisbin et al., 2012; Sohn and Xing, 2007; Sundquist et al., 2008; Bercovici et al., 2012; Baran et al., 2012). All of these approaches are based on a generative model of the admixed genome where the ancestral origin of each SNP (or a contiguous window) of the genome corresponds to the hidden states. These models specify the transition distribution between the ancestral states and the emission distribution of alleles at the genome conditioned on the ancestral state.

While such methods can be extended to infer local ancestries in ancient admixtures, we adopt a different perspective. Our method is based on the statistical framework of Conditional Random Field (CRFs). CRFs allow us to directly specify the conditional probability of the output state (*i.e.*, is this SNP or region derived from Neandertal) conditioned on a set of sufficient statistics computed from the data. The benefits of this framework is that we could choose sufficient statistics that are informative but have arbitrary dependence structure while in generative models such as HMMs, we would need to specify the complex joint distribution of these statistics. Further, statistical theory shows that discriminative models have lower asymptotic error than their generative counterparts, particularly when the model is misspecified (Ng and Jordan, 2002; Liang and Jordan, 2008). A final advantage of CRFs is that the likelihood can often be written as a convex function of the parameters which makes parameter estimation tractable. On the other hand, the likelihood of generative models is non-convex and parameter estimates obtained using the EM algorithm are not guaranteed to be the maximum likelihood estimates. A recent method (Maples et al., 2013) also uses a CRF for the problem of local ancestry inference in recently admixed populations and reports improved accuracies over a generative method (Baran et al., 2012).

A complementary strategy for detecting archaic admixture leverages patterns of variation in modern human genomes alone (Wall, 2000; Green et al., 2010). These methods are particularly useful for detecting admixture in the absence of a reference archaic genome (Hammer et al., 2011; Lachance et al., 2012). When a reference archaic genome is available, these methods have used the archaic genome to assess or amplify their accuracy (Wall et al., 2013; Green et al., 2010). An interesting direction for future work would be to compare maps built using these orthogonal approaches.

SI 1 Conditional Random Field for predicting archaic local ancestry

Consider m haplotypes in the test admixed population (say Europeans) $\{x_1, \ldots, x_m\}$ where each haplotype $x_s, s \in \{1, \ldots, m\}$ is a binary vector describing the allelic state at each of n SNPs, $x_s = (x_{s,1}, \ldots, x_{s,n})$. 0 and 1 denote the ancestral and derived alleles respectively (the determination of ancestral and derived allele state is explained later). Each haplotype is also associated with a random binary vector $Y_s = (Y_{s,1}, \ldots, Y_{s,n})$ where $Y_{s,j} = 1$ iff $x_{s,j}$ has Neandertal ancestry (more precisely, the lineage ancestral to $x_{s,j}$ passes through the Neandertal population during its history).

Our data also includes a panel of m_A African haplotypes, $h_t \in \{0,1\}^n, t \in \{1, \ldots, m_A\}$. From this panel, we estimate the derived allele frequency in Africans $\vec{p}_A = (p_{A,1}, \ldots, p_{A,n}), 0 \leq p_{A,j} \leq 1$. At each SNP, we also observe whether Neandertal carries the derived allele $\vec{p}_N = (p_{N,1}, \ldots, p_{N,n}), 0 \leq p_{N,j} \leq 1$. Using the high-coverage Neandertal genome, we use the diploid genotype calls to determine $p_{N,j} = 1$, if either of the two alleles is derived and 0 otherwise.

We aim to specify the distribution of the unobserved Neandertal ancestry vector Y_s given the observed genetic data. The conditional distribution of the ancestry vector y_s for haplotype s given data and parameters is specified by a Conditional Random Field (CRF) (Lafferty et al., 2001):

$$\begin{aligned}\log \Pr(y_s | x, h, s, \vec{r}, \vec{p}_A, \vec{p}_N, \vec{\alpha}, \vec{w}, \vec{\lambda}) &= \sum_{j=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{s,j}, x, h, \vec{p}_A, \vec{p}_N, s, j) \\ &+ \sum_{j=1}^{n-1} g(y_{s,j}, y_{s,j+1}, \vec{\lambda}, r_j) + \alpha_{y_1} - \log Z\end{aligned}$$

Here x is the $m \times n$ matrix of test haplotypes, h is the $m_A \times n$ matrix of African reference haplotypes. s is an index of the currently tested haplotype, $s \in \{1, \ldots, m\}$. $f_k, k \in \{1, \ldots, K\}$ and g are real-valued "feature" functions. f_k couples the unobserved ancestral state at a SNP allele of the test haplotype (*i.e.*, does this allele have Neandertal ancestry?) to the observed data and is analogous to emissions probabilities of a hidden Markov model (HMM). We will refer to f_k as emission functions and describe them in detail below. g couples the ancestral states at adjacent SNPs and is analogous to the transition probability matrix of a HMM. We will refer to g as the transition function. $\vec{\alpha} = (\alpha_0, \alpha_1), \alpha_0 + \alpha_1 = 1$ are related to the admixture proportions of the non-Neandertal and Neandertal ancestries respectively. w_k is the parameter associated with the emission function f_k. $\vec{\lambda}$ are parameters associated with the transition function g. r_j is the genetic distance between SNPs j and $j+1$ and Z is the log normalization constant so that the expression represents a valid probability distribution.

Of these quantities, x, h, $\vec{p}_A$ and $\vec{p}_N$ are data. $\vec{r}$ can be obtained from one of several fine-scale genetic maps (Hinch et al., 2011; Kong et al., 2010; Myers et al., 2005). While $\vec{\alpha}$ could also be estimated from the data, it makes a minor contribution to the likelihood and we instead fix this quantity at $(0.95, 0.05)$. The parameters to be estimated are $\vec{w}$ and $\vec{\lambda}$.

Given the parameters $\Theta = (\vec{w}, \vec{\lambda})$, and the non-African haplotypes x, we would like to infer the marginal probability that the SNP j of haplotype s is Neandertal, $\gamma_{s,j} = \Pr(y_{s,j} = 1 | x, \Theta)$ for $s \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$. As in the case of HMMs, these marginal probabilities can be computed efficiently using the forward-backward algorithm (Lafferty et al., 2001; Sutton and McCallum, 2011). Unlike a HMM however, the emission feature functions can have a more general conditional independence structure.

SI 1.1 Feature functions

We describe the emission functions, $f_k, k \in \{1, \ldots, K\}$. The emission functions relate the observed features in the data to the unobserved ancestral state $y_{s,j}$ at SNP j of haplotype s. We consider two classes of feature functions :

1. The first class summarizes the joint allelic configuration in Europeans, Africans and Neandertals at a single SNP.

   $$f_{1,(a,b,c)}(y_{s,j}, x, h, \vec{p}_A, \vec{p}_N, s, j) = \mathbf{1}\{y_{s,j} = 1 \& x_{s,j} = a \& p_{A,j} \in b \& p_{N,j} = c\}$$

   This class is a vector of features indexed by the numbers (a, b, c) which correspond to bins of the allelic configuration. Feature $f_{1,(a,b,c)}$ is 1 if the allelic state of the test haplotype s at SNP j, $x_{s,j}$, is a; if the allele frequency at this SNP in Africa falls in bin b; if the allele frequency at this SNP in Neandertals is in bin c; and if this allele in test haplotype s has Neandertal ancestry, *i.e.*, $y_{s,j} = 1$. In our application, $a \in \{0, 1\}$ corresponding to the test haplotype carrying a derived or ancestral allele at SNP s, $c \in \{0, 1\}$ corresponding to the Neandertal sequence carrying no or at least one derived allele at SNP s and $b \in \{0, 1, 2\}$ corresponding to the African derived allele frequency $p = 0$, $0 < p < 1$ or $p = 1$ respectively. Each of these features functions is associated with a parameter – corresponding to an element of $\vec{w}$. For example, if the parameter corresponding to the configuration $(a, b, c) = (1, 0, 1)$ is positive, this implies that the CRF is likely to assign Neandertal ancestry to the test haplotype at a SNP where this haplotype carries a derived allele that is absent in Africa but is carried by Neandertal. The strength of this preference is determined by the value of this parameter.

   In our application, we only used the features corresponding to the configurations $(1, 0, 1)$ and $(1, 1, 0)$ because these were found to be highly informative of Neandertal ancestry in simulations. In other words, the parameters associated with all the other features were set to 0. The configuration $(1, 0, 1)$ corresponds to a site at which the test haplotype carries a derived allele that is also found in the Neandertal sequence but is absent in the African sequences. Such a site has an increased likelihood of having Neandertal ancestry relative to a random site. The configuration $(1, 1, 0)$ corresponds to sites at which the test haplotype carries a derived allele that is observed to be polymorphic in the African samples but is absent in the Neandertal genome. Thus, under an infinite sites model, the local genealogy at this site consists of the lineage leading to test haplotype s coalescing with an African haplotype before coalescing with any of the Neandertal haplotypes. As a result, the test haplotype is less likely to carry Neandertal ancestry at this SNP.

   The binning of the allele frequency spectrum as well as the choice of informative features described here is arbitrary. It is likely that other ways of binning or other combinations of features might improve accuracy. We have not systematically tested these design choices in this work.

2. The second class of features compares the distance between the test haplotype and all other haplotypes in the African population to the distance between the test haplotype and the Neandertal sequence locally.

   $$f_2(y_{s,j}, x, b, \vec{p}_A, \vec{p}_N, s, j) = \mathbf{1}\left\{\frac{d_{j,L}(x_s, \vec{p}_N)}{min_{t \in \{1,\ldots,m_A\}} d_{j,L}(x_s, b_t)} < 1\right\}$$

   The specific function that we consider requires that the distance between the test haplotype and the Neandertal sequence be less than the minimum distance between the test and all other

African haplotypes. Here $d_{j,L}(x_s, b_t)$ denotes the distance between the test haplotype s and African haplotype t computed in non-overlapping sliding windows of length L that contains SNP j. While computing distance to Neandertal, we build a Neandertal haplotype haplotype $\tilde{x}$ from the frequency vector $\vec{p}_N$ by requiring that $\tilde{x}_j = \mathbf{1}\{p_{N,j} > 0\}$ so that $d_{j,L}(x_s, \vec{p}_N) = d_{j,L}(x_s, \tilde{x})$. A heterozygous site contributes the same as a homozygous site in computing distance. Thus, this distance is effectively the minimum distance of the potentially introgressed haplotype to the one of the two Neandertal haplotypes. We clarify this in the revised Methods. We set $L = 100$ Kb.

A number of improvements to the feature functions can be readily incorporated. *e.g.*, feature function f_1 can be extended to use the joint frequency spectrum of other modern human populations as well as other archaic groups such as the Denisovans. Further, the CRF is a meta-model. Any predictor that is informative of Neandertal ancestry *e.g*, posterior probabilities from other haplotype models, *e.g.*, (Price et al., 2009), could be included as a feature function.

The transition feature function g controls the transition probability across the hidden ancestry states. We define the transition feature function as follows:

$$g(y_{s,j}, y_{s,j+1}, \lambda, r_j) = \begin{cases} \lambda_{y_{s_j}, y_{s_{j+1}}} r_j, & y_{s,j+1} = y_{s,j} \\ \lambda_{y_{s_j}, y_{s_{j+1}}} + \log(r_j), & y_{s,j+1} \neq y_{s,j} \end{cases}$$

These transition functions are a convex approximation to the log transition probabilities of a standard Markov process modeling admixture between two populations with mixture proportions (α_0, α_1) and time of admixture t. In such a model, consider the log transition probability between states $(y_{s,j+1} = y_{s,j} = 0)$. For small genetic distances, this can be approximated as:

$$\begin{aligned} \log \Pr(y_{s,j+1}|y_{s,j}) &= \log\left((1 - \exp(-tr_j))\,\alpha_0 + \exp(-tr_j)\right), y_{s,j+1} = y_{s,j} = 0 \\ &\approx -(1-\alpha_0)tr_j \end{aligned}$$

Similarly, we can approximate:

$$\begin{aligned} \log \Pr(y_{s,j+1}|y_{s,j}) &\approx \log(t\alpha_1) + \log(r_j), & y_{s,j+1} = 1, y_{s,j} = 0 \\ \log \Pr(y_{s,j+1}|y_{s,j}) &\approx \log(t\alpha_0) + \log(r_j), & y_{s,j+1} = 0, y_{s,j} = 1 \\ \log \Pr(y_{s,j+1}|y_{s,j}) &\approx -(1-\alpha_1)tr_j, & y_{s,j+1} = 1, y_{s,j} = 1 \end{aligned}$$

The g feature function is obtained by replacing the parameters (α_0, α_1, t) by an unconstrained set $\vec{\lambda} = (\lambda_{ij}), i, j \in \{0, 1\}$. This convex approximation makes parameter estimation efficient.

SI 1.2 Parameter Estimation

To infer Neandertal ancestry using the CRF, we first need to estimate its parameters Θ. To estimate the parameters of the CRF, we need training data ,*i.e.*, haplotypes labeled with true Neandertal ancestries. In practice, we do not have access to training data. To get around this problem, we estimate the parameters of the CRF on data simulated under an appropriate demographic model. For simulated data, the ancestries of the haplotypes are known and can be used for parameter estimation. The CRF, with parameters estimated from simulations, is then be used to make inferences on real data.

Assume the training data consists of L unlinked loci. Each locus l has m haplotypes over $n_l, l = 1, \ldots, L$ SNPs. We denote the sequence of haplotypes at locus l by a matrix x^l. The true ancestral states of each SNP in every haplotype, represented by another $m \times n_l$ matrix o^l, are also known. We then estimate Θ that maximizes the regularized conditional log likelihood:

$$l(\Theta) = \sum_{l=1}^{L} \sum_{s=1}^{m} \frac{1}{m} \log \Pr({o^l}_s | x^l, s, \Theta) - \nu {\|\Theta\|_2}^2 \tag{1}$$

We regularize the conditional log likelihood using a L2 penalty to ensure that the optimization problem is strongly convex and to encourage parameter shrinkage. We maximize Equation 1 using a limited-memory version of LBFGS (Byrd et al., 1994), as implemented in the ALGLIB library (Bochkanov and Bystritsky). We set $\nu = 10$ although a broad range of values appear to work well in practice.

SI 2 Validation of the CRF

To assess the accuracy of the CRF for inferring Neandertal ancestry, we estimate the parameters of the CRF on simulated data and measure its precision and recall (defined below) on additional simulated data that was not used for parameter estimation. A concern, however, is that the inferences of the CRF might be sensitive to the demographic model used for parameter estimation. This is problematic as the true demographic model is only vaguely known in practice. To assess the effect of uncertainties in the demographic model, we evaluated the sensitivity of the inferences made by the CRF when we systematically perturbed the parameters of the demographic model.

SI 2.1 Simulations

We assumed a simple demographic model relating Africans, Europeans and Neandertals (see Figure SI 2.1). The parameters of the model were broadly constrained by the observed allele frequency differentiation F_{ST} between Africans and Europeans (which we measured as the F_{ST} between the HapMap YRI and CEU populations) and to the $D - statistics$, which measure the excess rate of sharing of alleles, between Europeans and Africans relative to Neandertals. On the simulated data, we measured $F_{ST} = 0.14$ and $D - statistics, D(Y, E; N, C) = 0.052$ (see Table SI 2.1 for the set of key parameters). We caution, however, that these statistics only loosely constrain the model. We comment on several aspects of this model:

1. The time of split between Neandertals and modern humans was set to 13000 generations. This is on the upper end of the estimates of the population split time of Neandertals and modern humans (Meyer et al., 2012). We chose this estimate to obtain a match to the D-statistic for a fixed gene flow proportion. However, we also assume that the Neandertal effective populations was constant and equal to 10000 throughout its history. This assumption is also at odds with observations from incomplete lineage sorting statistics from various Neandertals (Reich et al., 2010), PSMC curves (Prüfer et al., 2013), and from heterozygosity estimates (Prüfer et al., 2013), all of which indicate that Neandertals have had a reduced effective population size for an extended period of time. Assuming a more recent population split might need to be offset by modeling a reduced effective population size to match the D-statistic. We explore the robustness of the CRF to the split time in Section SI 2.3.

2. We model a modest bottleneck, with inbreeding coefficient 0.1, in the European population that predates Neandertal gene flow. This bottleneck corresponds to the bottleneck associated with the out-of-Africa event although the specific parameters associated with this bottleneck vary Keinan et al. (2007); Gutenkunst et al. (2009); Harris and Nielsen (2013). We assume a constant effective population size of 10000 in the European population after gene flow. Yang et al. (2012) show that a strong bottleneck after Neandertal gene flow does not match the observed doubly-conditioned site frequency spectrum. We again explore the robustness of the CRF to assumptions about the time as well as the strength of the bottleneck in Section SI 2.3.

3. Our model also does not include any recent gene flow between Europeans and Africans. Several studies (Wall et al., 2009; Harris and Nielsen, 2013) have suggested substantial gene flow between the two populations. It is clear that, under such models, the power of the CRF to infer Neandertal ancestry will be substantially reduced. A concern, however, is whether such a demographic scenario can lead to a high false discovery rate. We explore this issue in Section SI 2.3.

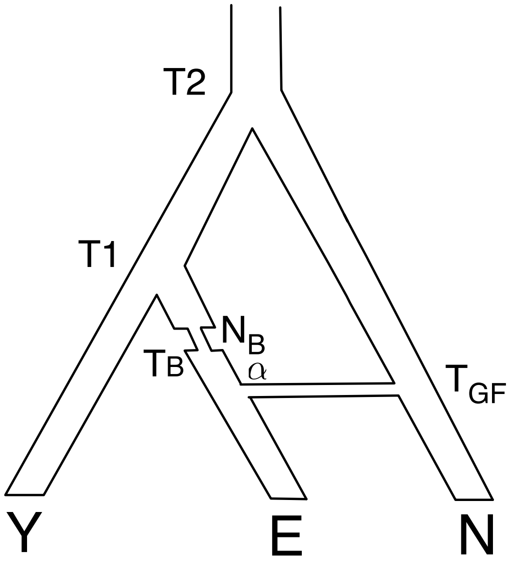


Figure SI 2.1: Basic demographic model used for parameter estimation

Further, our model assumes gene flow from Neandertals into modern Europeans occurred 1900 generations ago with Neandertals contributing about 3% genetic material, consistent with previous estimates (Green et al., 2010; Sankararaman et al., 2012). The gene flow was set to occur after the out-of-Africa bottleneck which is consistent with the observation in Yang et al. (2010)).

We simulated 120 European chromosomes, 118 African chromosomes and 1 Neandertal chromosome across 200 1Mb loci using the program msHOT (Hellenthal and Stephens, 2007). msHOT allows us to simulate data under a model that allows recombination hotspots. We chose parameters for the recombination model similar to the parameters described in Hellenthal et al. (2008). We modified msHOT to allow us to annotate regions of Neandertal ancestry in each simulated haplotype. We used 100 loci for parameter estimation (training data) and used the remaining 100 for assessing accuracy (test data). To ensure that our results could be extrapolated to real data, we also attempted to preprocess the simulated data to match the characteristics of the 1000 Genomes

T_1	Africa-Europe split	2500
T_2	Modern human-Neandertal split	13000
T_{GF}	Neandertal gene flow	1900
α	Neandertal admixture fraction	0.03
T_B	Time of bottleneck in Europe (end)	2400
N_B	Effective population size during European bottleneck	100
l_B	Duration of European bottleneck	20
μ	Mutation rate	2.5×10^{-8}

Table SI 2.1: Key parameters of the demographic model used for parameter estimation (see Figure SI 2.1).

data.

SNP calling in the 1000 Genomes project has low power to detect low-frequency variants (The 1000 Genomes Project Consortium, 2012). We simulated this process in our data by retaining SNPs based on their minor allele counts. SNPs with minor allele counts of $1, 2, 3, 4, 5, 6, 7, 8, \geq 9$ were accepted with probabilities 0.25, 0.50, 0.75, 0.80, 0.90, 0.95, 0.96, 0.97, 0.98, and 0.99 respectively. Further, we only retained positions that were polymorphic in the simulated European chromosomes. The resulting simulated dataset has an average SNP density of 3.3 per kb. The 1000 Genomes Phase I data, processed as described in Section SI 3, has an average density of 2.4 SNPs per kb.

We used the true genetic map as input to the CRF. Although we expect current genetic maps to be fairly accurate at large size scales, the maps are noisy at smaller scales (Hinch et al., 2011; Kong et al., 2010; Sankararaman et al., 2012). We also assume that the true haplotype phase is known. We assessed the robustness of our results to these assumptions below.

SI 2.2 Results

We evaluated the accuracy of the CRF to predict Neandertal ancestry at each allele in a haplotype in the test data. To do this, we declared an allele in a haplotype to be Neandertal if the marginal probability at the allele exceeded a threshold t: $\hat{o}^l_{s,j}(t) = \mathbf{1}\{{\gamma^l}_{s,j} \geq t\}$, $t \in [0, 1]$. We then compared these predictions to the true Neandertal ancestries. We varied t over the interval $[0, 1]$ in steps of $\frac{1}{1000}$. At each value of t, we compute the precision (fraction of predictions that are truly Neandertal) and the recall (fraction of truly Neandertal alleles that are predicted), defined formally as:

$$\begin{aligned}
Precision(t) &= \frac{TP(t)}{TP(t) + FP(t)} \\
Recall(t) &= \frac{TP(t)}{TP(t) + FN(t)} \\
TP(t) &= \sum_{l \in \{test\ data\}} \sum_{s=1}^{m} \sum_{j=1}^{n} \mathbf{1}\{\hat{o}^l_{s,j}(t) = 1 \& {o^l}_{s,j} = 1\} \\
FP(t) &= \sum_{l \in \{test\ data\}} \sum_{s=1}^{m} \sum_{j=1}^{n} \mathbf{1}\{\hat{o}^l_{s,j}(t) = 1 \& {o^l}_{s,j} = 0\} \\
FN(t) &= \sum_{l \in \{test\ data\}} \sum_{s=1}^{m} \sum_{j=1}^{n} \mathbf{1}\{\hat{o}^l_{s,j}(t) = 0 \& {o^l}_{s,j} = 1\}
\end{aligned}$$

Figure SI 2.2 shows a precision-recall curve of the CRF. The CRF attains a recall of around 15% at a precision of 99% and a recall of about 38% at a precision of 90%. At a threshold $t = 0.90$, the precision is 98% while the recall is 23%. For comparison, we also show the precision-recall curve for a "random" method that outputs the probability of Neandertal ancestry as a uniform number in $[0, 1]$. The results in Figure SI 2.2 show that the CRF is able to predict Neandertal ancestry with reasonable sensitivity while still being precise. There are several caveats associated with these results. Firstly, the CRF was tested on data generated from the same demographic model that was also used to generate training data. This assumption is unrealistic. It is challenging to assess accuracy over a comprehensive range of demographic models. Instead, we assessed the robustness of the CRF, *i.e.* whether it is likely to produce an elevated number of false positives when the true demographic model differs from the demographic model used for parameter estimation (Section SI 2.3). Secondly, in practice, errors in the Neandertal, Chimpanzee and modern human genomes could also reduce the reported accuracy. While we expect the model to be robust to random, uncorrelated errors, it is unclear how systematic or correlated errors might affect inference. Finally, we have also assumed that the true genetic map and the true phase are known. We now assess the impact of this assumption on the accuracy.

To assess the sensitivity of the results reported in Figure SI 2.2 to errors in the genetic map and to errors due to phasing, we paired the simulated haplotypes in the test data to form genotypes which we then phased with Beagle using default parameters (Browning and Browning, 2007). Further, we assumed that the genetic map is accurate at a 1Mb scale but is simply an extrapolation of the physical map at smaller scales so that the resulting map lacks hotspots. We applied the CRF to this perturbed data with parameters estimated on training data using the true genetic map and the true phase. Figure SI 2.2 shows that the lack of true phase and errors in the map leads to a decrease in accuracy, although the effect is small. At 90% precision, the recall is now 31% while at 99% precision, the recall is about 7%. At a threshold $t = 0.90$, the precision is 94% while the recall is 26%. These results indicate that the CRF, although trained on data where phase and genetic map is known, is quite robust to uncertainty in both. A possible explanation for this result is that Neandertal haplotypes are quite distinct and hence, easier to phase than a typical modern human haplotype particularly when the Neandertal haplotype is present in more than one copy in the sample.

SI 2.3 Robustness analysis

We were concerned that the procedure used for parameter estimation in the CRF makes it sensitive to the demographic model assumed. To assess the robustness of the CRF to misspecification of the demographic model, we followed a procedure similar to the one outlined in Lachance et al. (2012). We started with the demographic model described in Section SI 2.1 and perturbed each of its parameters in turn. For each of these perturbed parameters, we varied the proportion of Neandertal gene flow from 0.01 to 0.04. We did not attempt to match statistics of data simulated under each of these demographic models to the empirical values. We then applied the CRF, with parameters estimated under the original unperturbed demographic model, to each perturbed demographic model and assessed its false discovery rate.

Under each perturbed demographic model, we simulated 100 1 Mb regions using a version of ms which we modified to allow us to annotate regions of Neandertal ancestry (Hudson, 2002). We did not use msHOT for these experiments due to computational considerations. We simulated 100 chromosomes each from a European and African population and one chromosome from the Neandertals. We used a constant recombination rate of 1.3×10^{-8} per bp per generation for these simulations. We varied each of the following parameters of the demographic model :

1. the sequence mutation rate μ
2. T_2 – the time of split of modern humans and Neandertals
3. T_1 – the time of split of Africans and Europeans
4. T_{GF} – the time of Neandertal gene flow
5. the time T_B and the duration
6. l_B of the bottleneck in the European population
7. n_1 – the effective population size in the European population since gene flow

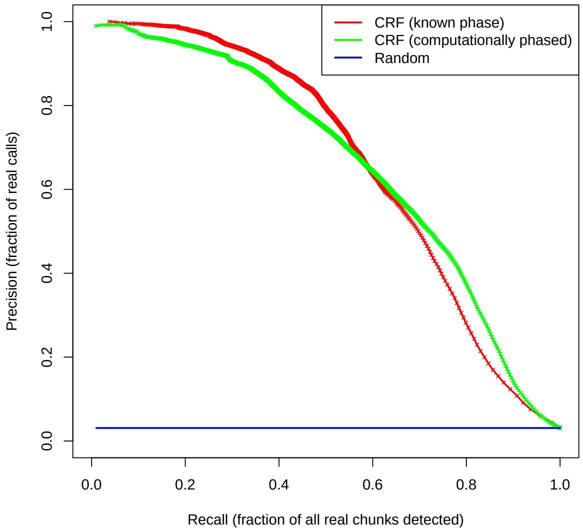


Figure SI 2.2: Precision-Recall curve for inferring Neandertal local ancestry using a high-coverage archaic genome.

We also assumed a bottleneck in the Neandertals beginning 6120 generations ago and ending 6000 generations ago, in which the Neandertal effective population size is reduced to 100. Without this bottleneck and assuming a constant effective population size of 10000 in Neandertals leads to a maximum $D - statistics, D(Africans, Europeans; Neandertal, Chimp) = 0.0278$ across all parameter settings, which is outside the confidence interval for the $D{-}statistic, D(Yoruba, French; Neandertal, Chimp) = 0.048 \pm 0.0059$ (Prüfer et al., 2013). It is plausible that the Neandertal population experienced a long-term reduction in its effective population size which we did not explicitly model here (Prüfer et al., 2013).

We also considered a demographic model proposed by Wall et al. (2009) – this model includes recent migration between Africans and Europeans as well as recent population growth in Europe, neither of which was part of the original demographic model used for parameter estimation.

We evaluated the false discovery rate ($= 1 - Precision$) when we restrict to sites at which the CRF assigns a marginal probability ≥ 0.9, *i.e.*, $\{(s, j) : \hat{o}^{l}_{s,j}(0.9) = 1\}$. Figure SI 2.3 shows that for all the demographic parameters, the false discovery rate is less than 0.1 suggesting that the model probabilities are conservatively calibrated. In fact, in about 97% of these simulations, the false discovery rate is less than 0.02.

We also computed another measure of accuracy. We defined Neandertal haplotypes by scanning each individual for a run of probabilities ≥ 0.9. We then declared a region as a false positive if it showed no overlap with a true Neandertal haplotype. Figure SI2.4 shows that the false discovery rate is still quite low, always well below the threshold of 0.1. 67% of the simulations have a false discovery rate < 0.01 (estimate + 1.96× se). When we restrict our analysis to Neandertal haplotypes that are at least 0.02 cM long, the accuracy increases further (Figure SI 2.5. 92% of the simulations have a false discovery rate < 0.01).

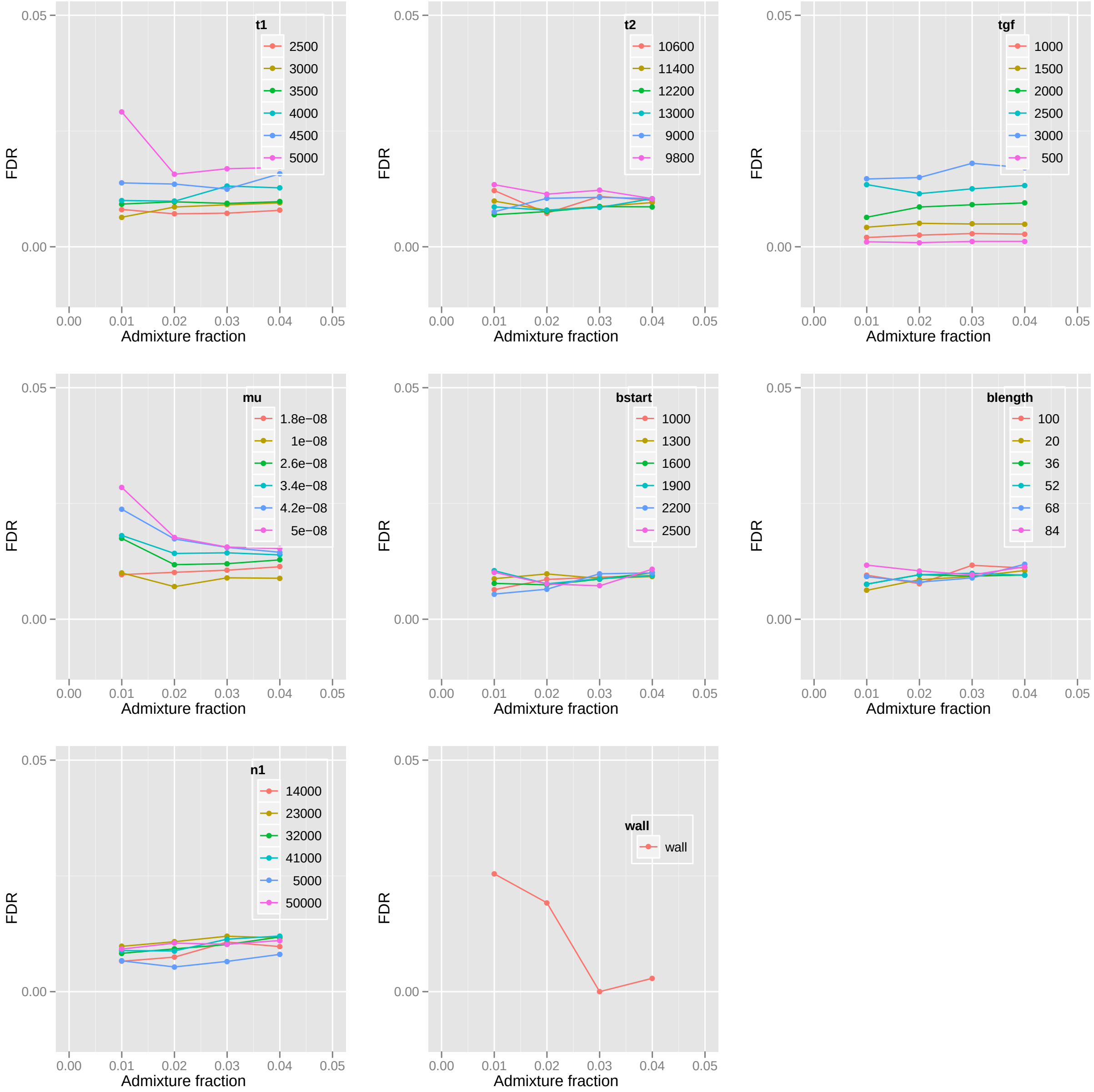


Figure SI 2.3: False discovery rates ($1 - Precision$) of the CRF for different demographic models. We varied each parameter of the basic demographic model described in Section SI 2.1. For each of these parameter settings, we varied the proportion of Neandertal gene flow from 0.01 to 0.04. We simulated data under each of these parameter settings. We estimated the false discovery rate when we restrict to sites with predicted probability ≥ 0.9. We do not show error bars as the estimated false discovery rates are quite precise (standard errors $< 10^{-4}$). The false discovery rate is less than 0.1 at this threshold across parameter settings. In about 97% of these simulations, the false discovery rate is less than 0.02.

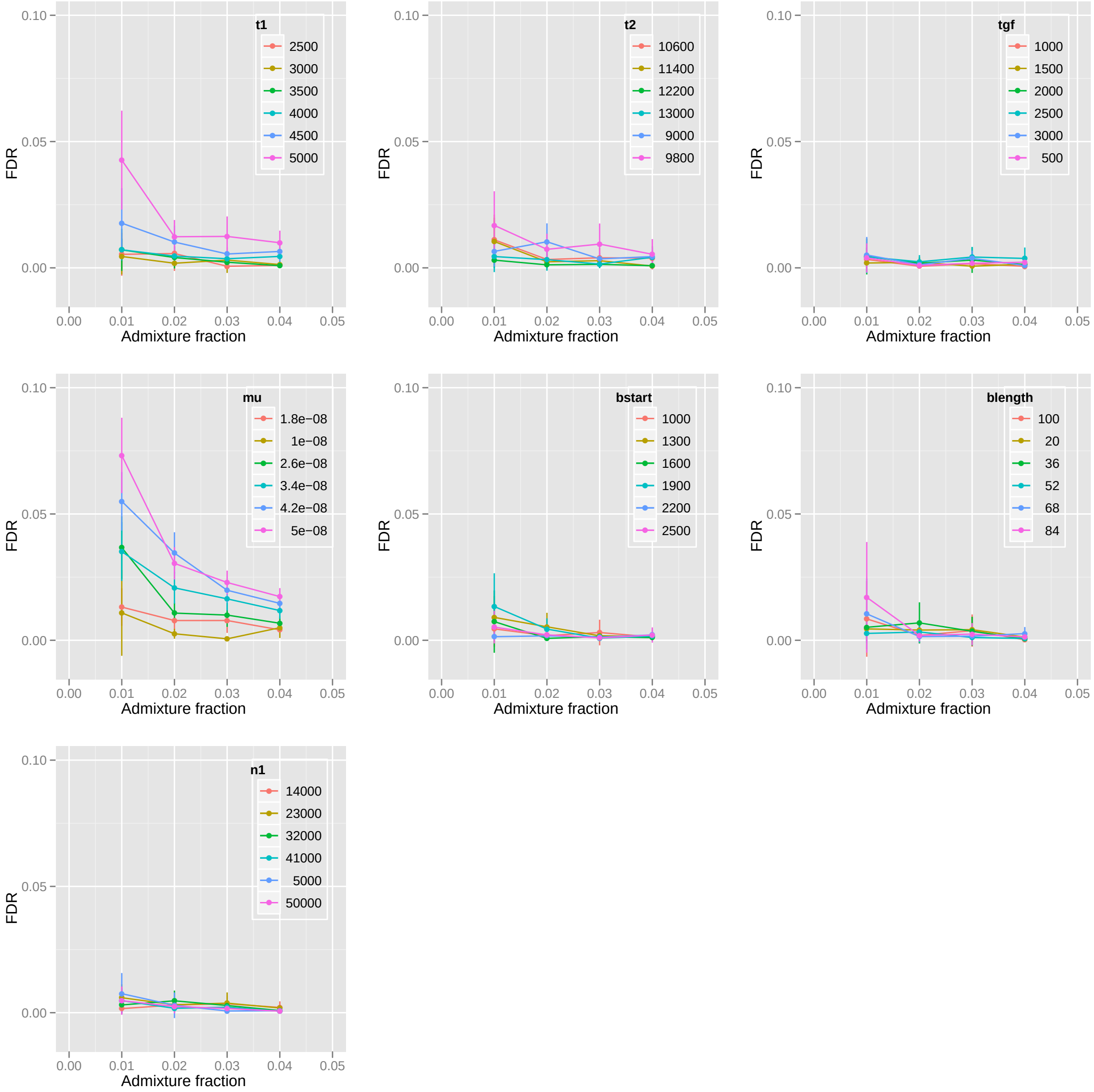


Figure SI 2.4: False discovery rates of the CRF for different demographic models. We varied each parameter of the basic demographic model described in Section SI 2.1. For each of these parameter settings, we varied the proportion of Neandertal gene flow from 0.01 to 0.04. We simulated data under each of these parameter settings. We defined Neandertal haplotypes as runs of alleles in each individual with marginal probability of Neandertal ancestry ≥ 0.9. We then declared a region as a false positive if it showed no overlap with a true Neandertal haplotype. We plot the estimate of the false discovery rate and 1.96× the standard error of this estimate (using a block jackknife with 100 blocks). The false discovery rate is always well below 0.1 and 66% of the simulations have a false discovery rate < 0.01.

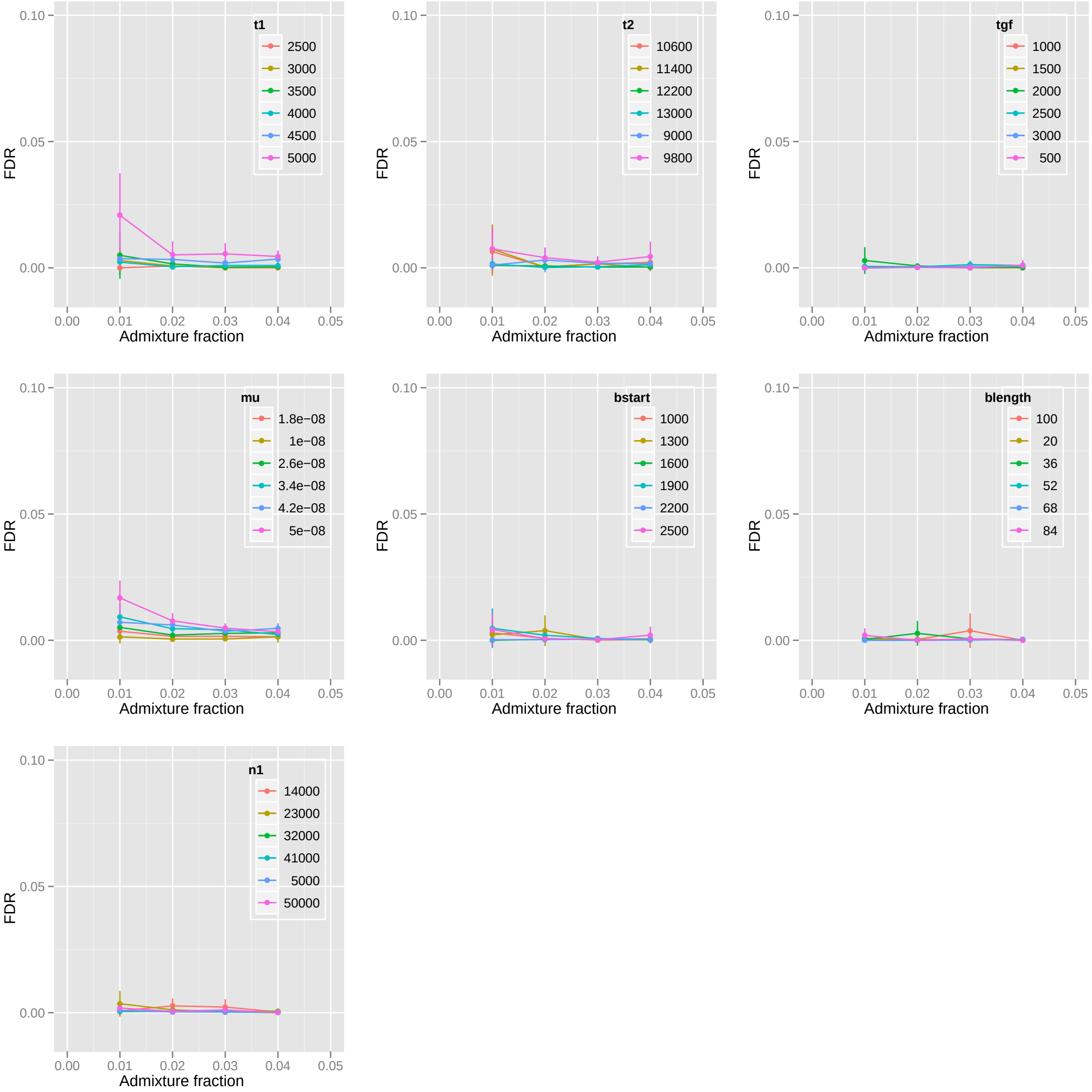


Figure SI 2.5: False discovery rates of the CRF for different demographic models. We varied each parameter of the basic demographic model described in Section SI 2.1. For each of these parameter settings, we varied the proportion of Neandertal gene flow from 0.01 to 0.04. We simulated data under each of these parameter settings. We defined Neandertal haplotypes as run of alleles in an individual with marginal probability of Neandertal ancestry ≥ 0.9 with genetic length > 0.02 cM. We then declared a region as a false positive if it showed no overlap with a true Neandertal haplotype. We plot the estimate of the false discovery rate and 1.96× the standard error of this estimate (using a block jackknife with 100 blocks). The false discovery rate is always well below 0.1 and 92% of the simulations have a false discovery rate < 0.01.

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

Table SI 2.2: Power to infer Neandertal ancestry as a function of the effective population size. *Recall* is computed at a precision of 90%. Standard errors were estimated by a block jackknife with 100 blocks.

	Recall
Autosomes	0.384±0.006
X	0.495±0.009

Table SI 2.3: Power to infer Neandertal ancestry on the X vs autosomes. *Recall* is computed at a precision of 90%. Standard errors were estimated by a block jackknife with 100 blocks.

SI 2.4 Power as a function of demographic and genomic features

We expect the power to infer Neandertal ancestry to depend on a number of parameters. One parameter that is expected to affect the power is the local effective population size. For example, the effective population size is known to vary along the genome and is reduced in regions with strong background selection and on the X chromosome. We expect that the power to infer Neandertal ancestry is increased in regions of reduced effective population size because the gene trees in humans are shorter in these regions so that Neandertal introgressed regions stand out more clearly.

To test this intuition, we used the default demographic parameters used in the Section SI 2.3. We varied the effective population size across all populations from an initial value of $10,000$ to $2,500$. We assessed the power to infer Neandertal ancestry at a precision of 90%. While the absolute estimates of power will depend on several demographic parameters, we see a clear trend of an increase in power when the effective population size is reduced (Table SI 2.2). A caveat of this analysis when used to interpret the effects of background selection is that background selection not only affects the mean coalescent time for a pair of alleles but also the shape of genealogies so that changing the effective population size does not fully capture the effects of background selection (Charlesworth et al., 1993, 1995; Williamson and Orive, 2002).

Further, to test the power of the CRF on the X chromosome, we simulated data with an effective population set to $\frac{3}{4}$ of the autosomal effective population size *i.e.*, $N_e = 7,500$. We also set the recombination rate to $\frac{2}{3}$ of the autosomal recombination rate (0.8667×10^{-8}) and the mutation rate to 0.87 times the autosomal mutation rate (Scally et al., 2012) (1.044×10^{-8}). We see that power is increased on chromosome X relative to the autosomes (Table SI 2.3).

SI 3 Results on the 1000 Genomes data

We applied the CRF to the computationally phased haplotypes in each of the 13 populations in the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) (excluding the west African Yoruba YRI). The parameters of the CRF were estimated on data from the demographic model (Table SI 2.1) as described in Section SI 2.1.

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 (Prüfer et al., 2013). We restricted our analysis to sites which passed the filters described in Prüfer et al. (2013) and for which GQ ≥ 30. These filters discard sites which are identified as repeats by the Tandem Repeat Finder (trf) or which have Phred-scaled $MQ < 30$, or which map to regions where the alignment is ambiguous or which fall within the upper or lower 2.5^{th} 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 more liberal $map35_{50\%}$ filter that requires that 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 samples. For each haplotype analyzed, we also restricted to the set of polymorphic sites in the population containing the haplotype. The ancestral allele was determined from the 6-primate EPO alignment and we further restricted our analysis to sites with confidently called ancestral alleles (Paten et al., 2008). After filtering, we obtained 26,493,206 SNPs on the autosomes and 817447 SNPs on chromosome X. Genetic distances were obtained from the combined LD map (Myers et al., 2005) lifted over to hg19 coordinates. For the X chromosome, we obtained a sex-averaged map by scaling the X chromosome LD-based map by $\frac{2}{3}$.

SI 3.1 Plausibility of the map of Neandertal ancestry

We applied the CRF to each of the thirteen 1000 Genomes populations, except the YRI which we used as one of our reference populations. As a test of the robustness of the model, we also applied the CRF to populations with substantial African ancestry *i.e.*, LWK and ASW. In this test of robustness, we used the same parameters for the LWK and ASW as for the non-African populations even though we know that the non-African populations share more genetic material with Neandertals than the African and these populations do not share the Neandertal gene flow event common to the non-African populations (Green et al., 2010). We also present several results by combining inferences across populations. For example, we combined results across the CEU, GBR, FIN, IBS and TSI populations to obtain results for the European populations (denoted EUR). Analogously, we obtained results for East Asians (ASN consisting of CHB, CHS, JPT) and Americans (AMR consisting of CLM, MXL and PUR). For these combined analyses, we applied the CRF to each individual with a subpopulations and then averaged the results across all individuals that belong to a given population.

We defined *Neandertal haplotypes* by scanning for runs of consecutive alleles along a haplotype with a marginal probability > 0.9. Discarding haplotypes with zero physical or genetic lengths (*e.g.*, single SNPs predicted to have Neandertal ancestry), we predict tens of thousands of Neandertal haplotypes in each of these populations SI 3.1.

SI 3.1.1 Gross features of the predicted Neandertal ancestry

We analyzed several gross features of the predicted Neandertal ancestry. To do so, we estimated the proportion of the genome that is confidently inferred to be Neandertal, $tia(s)$, as the fraction of sites for which the marginal probability > 0.9.

$$tia(s) = \frac{1}{|H(s)|} \sum_{t \in H(s)} \frac{\sum_{j=1}^{n} \mathbf{1}\{\gamma_{t,j} > 0.9\}}{n} \tag{2}$$

Here $H(s)$ indexes the haplotypes that belong to individual s.

The above equation also holds for estimating Neandertal ancestry on the X chromosome. In the case of the X chromosome, we average over both chromosomes for females only.

Tables SI 3.2 and SI 3.3 lists the distribution of Neandertal ancestry proportions across the populations. The proportion of the genome that is determined to be confidently Neandertal across Eurasian populations ranges from 0.99 to 1.54. In the Luhya (LWK), the proportion of the genome confidently inferred to be Neandertal is 0.08%, an order of magnitude smaller than in non-Africans as expected from a population that carries little or no Neandertal ancestry (Table SI3.2) (a caveat is that some of the inferred lower Neandertal ancestry in the LWK could be due to their relatedness to the YRI who we use to screen out modern human alleles). The proportion of the genome confidently inferred to be Neandertal has a mean of 1.38% in East Asians and 1.15% in Europeans (Table SI 3.3), consistent with more Neandertal ancestry in East Asians than in Europeans (Meyer et al., 2012; Wall et al., 2013).

The standard deviation in the proportion of Neandertal ancestry across individuals from the same Eurasian population is $0.06 - 0.10\%$. To compute the theoretical value of this standard deviation, we assume that the admixture occurred in a single generation (a pulse model) followed by random mating for $T = 2000$ generations with a Neandertal admixture proportion of 2%. We ignore the effects of drift since admixture. The total genetic length is calculated to be 26.39 Morgans using a recombination rate of 1.3 cM/Mb and a genome size of 2.03 Gb based on the number of bases of the high-coverage Neandertal genome that pass the filters described in Prüfer et al. (2013) (we use the $map35_{50\%}$ mappability filter to compute this number as this is the filter that we use for the CRF). Gravel (2012, Equation 8) (ignoring the effect of drift) shows that the standard deviation under these parameters is 0.06%. In Section SI 5, we estimate the drift since Neandertal gene flow in Europeans and East Asians to be ≈ 0.10. Assuming that this drift corresponds to a constant effective population size and ignoring the effect of drift on the lengths of ancestry switch points, the standard deviation is 5.8% (Gravel, 2012, Equation 8).

SI 3.1.2 Comparison of predictions with the Vindija Neandertal

One concern with the inferred Neandertal ancestry is that contamination and errors in the Neandertal sequence might bias these estimates. Contamination at the read level has been estimated to be less than 2%. The high-coverage of the Neandertal genome implies that the genotypes are likely to be very accurate.

Nevertheless, to assess the robustness of our inferences to such effects, we estimated Neandertal ancestries in the 1000 Genomes Project CEU population using as Neandertal reference the low-coverage draft Neandertal genome (Green et al., 2010) instead of the high-coverage Neandertal genome. The draft genome was generated using DNA obtained from 3 bones found in Vindija cave in Croatia while the high-coverage genome was generated using DNA from a bone found in Denisova cave in Siberia.

To use the Vindija genome, we restricted our analysis to sites in the 1000 Genomes Project data that have at least one overlapping Vindija Neandertal read that passed quality filters. Reads were required to have mapping quality scores between 60 and 90 and base quality scores of at least 40 (Green et al., 2010). The low coverage of the Vindija genome does not allow confident calling of genotypes. Note, however, that the features used by the CRF (see Section SI 1.1) only require the frequency of the derived allele in Neandertal. We estimate this as the fraction of reads at a site that carry the derived allele. We use this frequency to define $\vec{p}_N$.

To compare the estimates of Neandertal ancestry when we use either of the two Neandertal genomes, we consider non-overlapping 100 Kb windows. Within each window w, we estimated

- the fraction of confidently inferred Neandertal haplotypes, $ta_t(w)$,

  $$ta_t(w) = \frac{\sum_{j \in w} \sum_{s=1}^{m} \mathbf{1}\{\gamma_{s,j} > t\}}{m|\{j \in S(w)\}|}$$

  We set $t = 0.90$.

- the average Neandertal ancestry

  $$la(w) = \frac{\sum_{j \in S(w)} \sum_{s=1}^{m} \gamma_{s,j}}{m|\{j \in S(w)\}|}$$

  Here $S(w)$ refers to the set of SNPs that belong to window w.

We restrict our analyses to windows that contain at least 10 SNPs that pass filters. The Spearman's rank correlation coefficient between the Neandertal ancestries estimated using the high-coverage and Vindija genomes at a 100Kb size scale is 0.88 using the fraction of confident Neandertal haplotypes and 0.94 using the average Neandertal ancestry. Figure SI 3.1 shows the scatterplot of the ancestries within each 100 Kb window using Vindija or the high-coverage Altai genome. We see that most windows have concordant ancestries.

To get another view of the concordance, we restricted attention to Neandertal haplotypes, defined as a consecutive run of SNPs assigned marginal probability of at least 0.9. We then divided the genome into non-overlapping 100 Kb windows. We determined that a window is introgressed if at least one of the haplotypes within that window overlaps a Neandertal haplotype. We then estimated the concordance of this estimate of introgression when we use either Vindija or Altai. Table SI 3.7 shows that the estimates of introgression are concordant in 93% of the windows. We do see a larger number of windows for which we detect introgression only when the Altai genome is used which is expected given its better quality. The estimates of Neandertal ancestry are largely concordant across the two Neandertal genomes used and are unlikely to represent contamination or errors in the Neandertal sequence. The differences in the estimates across the Vindija and Altai genomes could be either due to errors or differential sequence quality. On the other hand, these differences could also reflect genuine differences in the Neandertal sequences as well as differential relatedness of the genealogies relating the sequenced Neandertal to the introgressing Neandertal.

SI 3.1.3 Empirical estimate of the accuracy

In Section SI 2.2, we validated the accuracy of the CRF using simulations. We also attempted to obtain an estimate of the empirical accuracy of the CRF. To estimate the accuracy of the CRF on 1000 Genomes data, we make several assumptions:

Populations	Number of individuals	Neandertal haplotypes		Neandertal haplotypes	
Populations		Autosomes		X chromosomes	
		All	(> 0.02 cM)	All	(> 0.02 cM)
CEU	85	48355	35593	304	211
FIN	93	54238	39925	317	231
GBR	89	50683	37313	309	231
IBS	14	7100	5391	53	38
TSI	98	54521	39877	418	273
CHB	97	66419	49292	452	364
CHS	100	67734	50267	402	319
JPT	89	59956	44615	299	258
CLM	60	35337	25712	183	145
MXL	66	41076	29648	222	174
PUR	55	29949	21661	187	122
LWK	97	4223	2474	52	38
ASW	61	11526	8048	75	54

Table SI 3.1: Number of confidently predicted Neandertal haplotypes in each of the 1000 Genomes populations.

Populations	Neandertal ancestry	Neandertal haplotypes		Haplotype length	
	tiaAll		(> 0.02 cM)	All	(> 0.02 cM)
CEU	1.17±0.08	569±35	419±26	0.07±0.09	0.09±0.10
FIN	1.20±0.07	583±28	429±20	0.07±0.08	0.09±0.09
GBR	1.15±0.08	569±38	419±25	0.07±0.09	0.09±0.10
IBS	1.07±0.06	507±28	385±23	0.07±0.09	0.09±0.09
TSI	1.11±0.07	556±30	407±21	0.07±0.09	0.09±0.10
CHB	1.40±0.08	685±33	508±24	0.07±0.09	0.09±0.09
CHS	1.37±0.08	677±36	503±25	0.07±0.09	0.09±0.09
JPT	1.38±0.10	674±41	501±28	0.07±0.09	0.09±0.09
CLM	1.14±0.12	589±63	429±47	0.07±0.08	0.09±0.09
MXL	1.22±0.09	622±47	449±32	0.07±0.08	0.09±0.09
PUR	1.05±0.12	545±64	394±47	0.07±0.09	0.09±0.09
LWK	0.08±0.02	44±9	26±6	0.04±0.05	0.07±0.06
ASW	0.34±0.22	189±125	132±88	0.06±0.08	0.09±0.09

Table SI 3.2: Summary of predicted Neandertal ancestry across the autosomes in 1000 Genomes populations. Thresholded Neandertal ancestry refers to the fraction of positions which have a posterior probability > 0.9 and is estimated by $tia(s)$. The table reports the number of Neandertal haplotypes in each population (defined in Section SI 3.1) as well as the number of haplotypes that are longer than 0.02 cM. The table also reports the mean and standard deviation of the lengths of these haplotypes.

Populations	Neandertal ancestry	Neandertal haplotypes		Haplotype length	
	tia	All	(> 0.02 cM)	All	(> 0.02 cM)
EUR	1.15±0.08	567±36	417±25	0.07±0.09	0.09±0.09
ASN	1.38±0.08	679±37	504±26	0.07±0.09	0.09±0.09
AMR	1.14±0.13	588±66	426±48	0.07±0.08	0.09±0.09

Table SI 3.3: Summary of predicted Neandertal ancestry across the autosomes in non-African continental populations. Thresholded Neandertal ancestry refers to the fraction of positions which have a posterior probability > 0.9 and is estimated by $tai(s)$. The table reports the number of Neandertal haplotypes in each population (defined in Section SI 3.1) as well as the number of haplotypes that are longer than 0.02 cM. The table also reports the mean and standard deviation of the lengths of these haplotypes.

Populations	Neandertal ancestry	Neandertal haplotypes		Haplotype length	
	tia	All	(> 0.02 cM)	All	(> 0.02 cM)
CEU	0.21±0.17	5±3	4±2	0.06±0.06	0.08±0.06
FIN	0.19±0.14	4±2	3±2	0.06±0.06	0.08±0.06
GBR	0.20±0.15	4±2	3±2	0.06±0.06	0.08±0.07
IBS	0.23±0.18	5±3	4±2	0.06±0.05	0.07±0.05
TSI	0.25±0.20	6±3	4±2	0.06±0.06	0.08±0.07
CHB	0.30±0.21	6±3	5±2	0.06±0.07	0.07±0.07
CHS	0.27±0.21	5±3	4±2	0.06±0.08	0.08±0.08
JPT	0.26±0.21	5±3	4±3	0.07±0.07	0.08±0.07
CLM	0.22±0.16	4±2	3±1	0.09±0.10	0.11±0.10
MXL	0.21±0.15	5±2	4±2	0.08±0.07	0.09±0.08
PUR	0.20±0.15	5±2	3±1	0.06±0.06	0.09±0.07
LWK	0.04±0.07	2±1	1±1	0.04±0.03	0.05±0.03
ASW	0.07±0.11	3±2	2±2	0.06±0.05	0.07±0.04

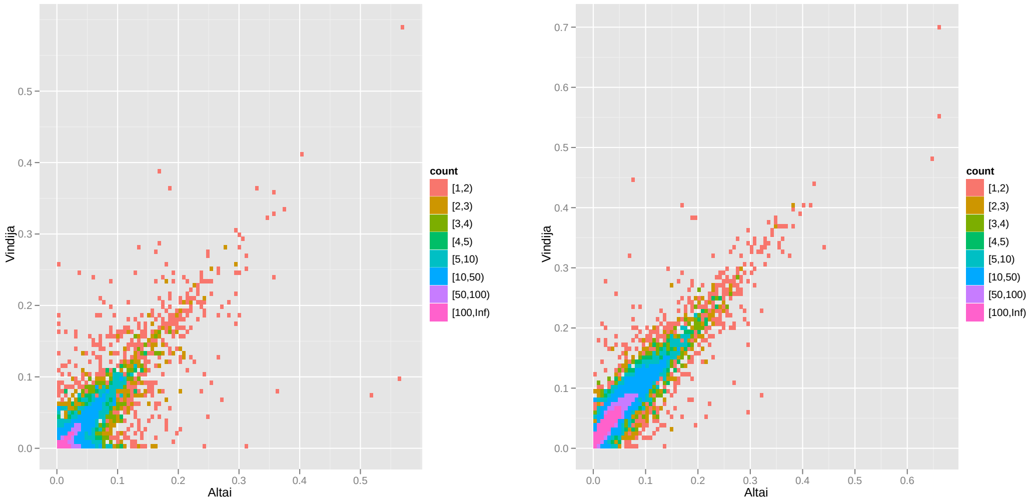
Table SI 3.4: Summary of predicted Neandertal ancestry on the X chromosome in the 1000 Genomes populations. Thresholded Neandertal ancestry refers to the fraction of positions which have a posterior probability > 0.9 and is estimated by $tia(s)$. The table reports the number of Neandertal haplotypes in each population (defined in Section SI 3.1) as well as the number of haplotypes that are longer than 0.02 cM. The table also reports the mean and standard deviation of the lengths of these haplotypes.

Populations	Neandertal ancestry	Neandertal haplotypes		Haplotype length	
	tia	All	(> 0.02 cM)	All	(> 0.02 cM)
EUR	0.21±0.17	5±3	4±2	0.06±0.06	0.08±0.06
ASN	0.28±0.21	5±3	4±2	0.06±0.07	0.08±0.07
AMR	0.21±0.15	4±2	3±2	0.08±0.08	0.10±0.08

Table SI 3.5: Summary of predicted Neandertal ancestry on the X chromosome in the 1000 Genomes non-African populations. Thresholded Neandertal ancestry refers to the fraction of positions which have a posterior probability > 0.9 and is estimated by $tia(s)$. The table reports the number of Neandertal haplotypes in each population (defined in Section SI 3.1) as well as the number of haplotypes that are longer than 0.02 cM. The table also reports the mean and standard deviation of the lengths of these haplotypes.

Chromosome	Neandertal ancestry		
	EUR	ASN	AMR
1	1.30±0.26	1.54±0.26	1.28±0.30
2	1.22±0.27	1.17±0.25	1.11±0.26
3	1.17±0.26	1.28±0.25	1.17±0.27
4	1.11±0.25	1.56±0.31	1.17±0.35
5	0.73±0.21	1.15±0.29	0.85±0.29
6	1.64±0.40	2.08±0.43	1.54±0.40
7	1.10±0.29	1.09±0.34	0.99±0.29
8	0.95±0.27	0.66±0.19	0.83±0.25
9	1.44±0.36	2.31±0.47	1.51±0.42
10	1.35±0.35	2.24±0.45	1.50±0.45
11	1.14±0.45	1.45±0.31	1.15±0.40
12	1.82±0.40	2.21±0.44	1.71±0.48
13	1.07±0.37	1.10±0.33	0.99±0.36
14	1.65±0.52	1.55±0.45	1.56±0.57
15	1.05±0.47	1.30±0.43	1.26±0.51
16	0.76±0.27	1.13±0.39	0.85±0.39
17	0.26±0.14	0.52±0.26	0.32±0.16
18	0.93±0.33	0.86±0.27	0.91±0.34
19	0.64±0.27	0.70±0.46	0.68±0.41
20	0.86±0.35	0.85±0.29	0.89±0.36
21	0.70±0.40	0.35±0.25	0.56±0.36
22	0.89±0.41	1.06±0.51	0.82±0.42
X	0.21±0.17	0.28±0.21	0.21±0.15

Table SI 3.6: Neandertal ancestry as estimated by $tia(s)$ stratified by chromosome.



(a)

(b)

Figure SI3.1: Comparison of the average Neandertal ancestry across the 1000 Genomes Project CEU individuals in 100 kb windows when we use either the high-coverage Altai or the Vindija Neandertal sequences. a) estimates the fraction of confident Neandertal alleles within each window , b) the average Neandertal ancestry.

	Altai	
Vindija	Not introgressed	Introgressed
Not introgressed	17367	1642
Introgressed	208	6872

Table SI 3.7: Concordance of Neandertal ancestry estimates when using either the high-coverage Altai or the Vindija Neandertal sequence.

- We assume that the African Luhya (LWK) have no Neandertal ancestry. Under this assumption, any Neandertal ancestry inferred in the LWK is a false positive. For a fixed threshold t to call an allele as Neandertal, denote the false discovery rate as $fp(t)$, *i.e.*, the fraction of alleles in LWK at which the marginal probability exceeds t.

- We assume that the false discovery rate in each non-African population tested is equal to the false discovery rate estimated in the LWK.

- We assume that the proportion of true Neandertal ancestry in the test non-African population is α. α has been estimated to be 0.0172 ± 00012 in Europeans and 0.0189 ± 0.0013 in Eastern non-Africans (Prüfer et al., 2013).

If the fraction of sites with a marginal probability of Neandertal ancestry of at least t in a tested non-African population is denoted $g(t)$, we can then estimate

$$\begin{aligned} Precision(t) &= \frac{g(t) - fp(t)}{g(t)} \\ Recall(t) &= \frac{g(t) - fp(t)}{\alpha} \end{aligned}$$

We vary t over the interval $[0, 1]$ in steps of $\frac{1}{100}$ to obtain an empirical precision recall curve. We analyzed the predicted Neandertal ancestry in Europeans and East Asians using the point estimates of α above. To plot the empirical precision-recall curve (Figure SI 3.2), we only retained those values of t which are not dominated by any other t, *i.e.*, there does not exist $s \neq t$ such that $Precision(s) > Precision(t)$ and $Recall(s) > Recall(t)$ or $Precision(s) = Precision(t)$ and $Recall(s) > Recall(t)$ or $Recall(s) = Recall(t)$ and $Precision(s) > Precision(t)$. Figure SI 3.2 shows that at a precision of 90%, the CRF attains a recall of 72% in Europeans and 85% in East Asians. At a probability threshold of 0.90, the CRF attains a recall of 62% at a precision of 93% in Europeans and a recall of 69% at a precision of 95% in East Asians. The recall estimate depends on the estimate of the total proportion of Neandertal ancestry α. The recall is lowest if the true Neandertal ancestry proportion is high. If we assume that the true α is 2 standard deviations above the point estimate, the recall at 90% precision is now 63% and 75% in Europeans and East Asians respectively, while at a threshold of 0.90, the recall is 55% and 61% in Europeans and East Asians respectively.

The high-coverage Neandertal genome has been observed to carry large regions of homozygosity consistent with recent inbreeding (Prüfer et al., 2013). Intuitively, we expect power to be lower in these regions because at these loci we have fewer Neandertal haplotypes making it more difficult to find the haplotypes that are related to the test introgressing sequence.

To test this idea, we used the tracts of homozygosity (HBD) longer than 2.5 cM identified in the high-coverage Neandertal genome (Prüfer et al., 2013). We divided the genome into non-overlapping windows of length w. We computed the fraction of confidently inferred Neandertal haplotypes, $ta(w)$ within each window.

$$ta_t(w) = \frac{\sum_{j \in w} \sum_{s=1}^{m} \mathbf{1}\{\gamma_{s,j} > t\}}{m|\{j \in S(w)\}|} \tag{3}$$

Here $S(w)$ refers to the set of SNPs that belong to window w. We chose a threshold $t = 0.90$ and $w = 1$ Mb. We restrict our analyses to windows that contain at least 10 SNPs that pass filters. We also restrict to windows that completely overlap or completely avoid an identified HBD tract. Note that it is not strictly necessary to analyze the relationship of Neandertal ancestry to HBD in

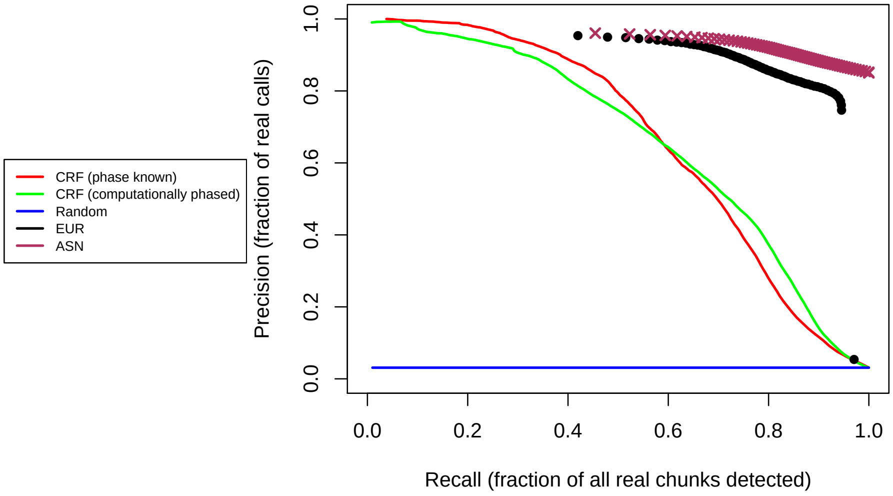


Figure SI 3.2: Precision-Recall curve for inferring Neandertal local ancestry using a high-coverage archaic genome. Curves are shown for the CRF on simulated data for two cases : i) where the genetic map and haplotype phase are known perfectly and ii) where the genetic map is accurate at a 1 Mb scale but does not capture recombination hotspots and the haplotype phase is estimated. Precision-recall curves for the CRF on European (EUR) and East Asian (ASN) populations in the 1000 Genomes dataset are also shown. These empirical curves were estimated assuming that any Neandertal ancestry detected in the sub-Saharan African population Luhya from Kenya (LWK) by the CRF is a false positive and that the method has the same false discovery rates in other non-African populations and using the estimates of the fraction of Neandertal ancestry in a population. As a baseline, we also simulated a random method, which assigns a random number uniformly distributed in $[0, 1]$, to predict Neandertal ancestry.

windows. However, this windowing analysis is useful to assess statistical significance using a block jackknife.

We find that the average proportion of the genome that is labeled as Neandertal with a confidence of >0.90, $ta_{0.90}(w)$, is 0.19% lower in large homozygous regions of the Neandertal genome than in the rest of the genome. To assess statistical significance, we performed a block jackknife, using 1 Mb blocks, of the difference of the ancestry estimates across the two regions. We use this procedure to obtain a jackknife estimate of the difference as well as a jackknife estimate of the standard error. We then use the estimator and the standard error to obtain a z-score which we convert to a two-sided P-value. We obtain a p-value=0.044. As expected, however, we do not see a significant difference between the Neandertal ancestry proportions in homozygous and non-homozygous regions of the genome measured using the unbiased estimate of ratios of S-statistics as $\frac{S(Eurasia,Africa;Denisova,Chimp)}{S(Neandertal,Africa;Denisova,Chimp)}$ proposed in Reich et al. (2010). (p-value=0.118). We also do not see a significant difference in power, estimated as the ratio of the first estimate to the second estimate, across the two regions (p-value=0.755). Repeating the analysis at $w = 10$ Mb, we see that the respective P-values are 0.951, 0.588 and 0.433 respectively.

SI 3.2 Variation of Neandertal ancestry along the genome

To assess variation in Neandertal ancestry along the genome, we computed the fraction of confidently inferred Neandertal haplotypes, $ta(w)$, within 100 Kb non-overlapping windows that tile each chromosome (or genome).

$$ta_t(w) = \frac{\sum_{j \in w} \sum_{s=1}^{m} \mathbf{1}\{\gamma_{s,j} > t\}}{m|\{j \in S(w)\}|} \tag{4}$$

Here $S(w)$ refers to the set of SNPs that belong to window w. We chose a threshold $t = 0.90$. We restrict our analyses to windows that contain at least 10 SNPs that pass filters. For each set of chromosomes (autosomes or chromosome X), we then estimated the Gini coefficient of the distribution of $ta_{0.90}$ of tiling windows (Gini, 1912). The Gini coefficient is a measure of the dispersion of Neandertal ancestry across the chromosome (rel). The Gini coefficient estimates the dispersion of Neandertal ancestry. Table SI3.8 shows the Gini coefficients on the autosomes and the X chromosome in the 1000 Genomes populations as well as the fraction of confidently inferred Neandertal haplotypes. To further analyze the estimates of this coefficient, we rank-ordered windows in decreasing order of the proportion of Neandertal ancestry (as measured by $ta_{0.9}$). For each rank, we then computed the cumulative Neandertal ancestry in windows with higher ranks as a fraction of the total Neandertal ancestry. We then computed the minimum fraction of windows needed to capture a given fraction f of Neandertal ancestry $f = \frac{i}{20}, i \in \{1, \ldots, 20\}$. If Neandertal ancestry were evenly distributed in the genome, a fraction f of the Neandertal ancestry would be found in a fraction f of windows. Table SI 3.9 shows this distribution for EUR and ASN and for the autosomes and chromosome X. Table SI 3.10 repeats this analysis using $ta_{0.25}$. We note that all the Neandertal ancestry is found within $35-50\%$ of the autosomes but within 20% of chromosome X. 95% of the Neandertal ancestry is found within $20-30\%$ of the autosomes but within 10% of chromosome X.

Populations	Individuals	Neandertal ancestry (%)		Gini coefficient(%)	
		Autosomes	X	Autosomes	X
CEU	85	1.17±0.08	0.21±0.17	82-94	96
FIN	93	1.20±0.07	0.19±0.14	81-94	97
GBR	89	1.15±0.08	0.20±0.15	83-94	97
IBS	14	1.07±0.06	0.23±0.18	86-96	98
TSI	98	1.11±0.07	0.25±0.20	82-94	97
CHB	97	1.40±0.08	0.30±0.21	82-94	97
CHS	100	1.37±0.08	0.27±0.21	83-96	97
JPT	89	1.38±0.10	0.26±0.21	82-95	97
CLM	60	1.14±0.12	0.22±0.16	80-93	97
MXL	66	1.22±0.09	0.21±0.15	82-93	97
PUR	55	1.05±0.12	0.20±0.15	80-92	97
LWK	97	0.08±0.02	0.04±0.07	91-97	98
ASW	61	0.34±0.22	0.07±0.11	82-93	97

Table SI 3.8: 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 (probability $> 90\%$) in each individual. The table reports the average and standard deviation of this measure across individuals within each population. We also report a measure of the variability (the Gini coefficient) of the fraction of confidently Neandertal alleles within non-overlapping 100 kb windows that tile the autosomes or chromosome X respectively. The Gini coefficient is zero when every window has the same ancestry and nearly 100 when all the Neandertal ancestry lies in one window but is zero elsewhere.

Neandertal	EUR		ASN	
ancestry(%)	Autosomes	Chromosome X	Autosomes	Chromosome X
5	0.002	0.001	0.002	0.001
10	0.004	0.001	0.004	0.001
15	0.007	0.003	0.006	0.002
20	0.011	0.004	0.010	0.004
25	0.015	0.005	0.013	0.004
30	0.019	0.006	0.017	0.005
35	0.024	0.008	0.022	0.006
40	0.030	0.009	0.027	0.008
45	0.037	0.011	0.033	0.008
50	0.045	0.013	0.040	0.010
55	0.054	0.015	0.047	0.012
60	0.064	0.018	0.056	0.013
65	0.075	0.020	0.065	0.015
70	0.088	0.024	0.076	0.018
75	0.104	0.029	0.089	0.022
80	0.123	0.034	0.105	0.027
85	0.147	0.042	0.125	0.032
90	0.179	0.052	0.152	0.039
95	0.228	0.066	0.193	0.051
100	0.414	0.113	0.351	0.086

Table SI 3.9: Fraction of 100 kb windows that contain $\frac{i}{20}, i \in \{1, \ldots, 20\}$ of the Neandertal ancestry as measured by $ta_{0.90}$.

Neandertal	EUR		ASN	
ancestry(%)	Autosomes	Chromosome X	Autosomes	Chromosome X
5	0.003	0.001	0.002	0.001
10	0.006	0.003	0.005	0.001
15	0.010	0.004	0.009	0.002
20	0.015	0.006	0.013	0.002
25	0.020	0.007	0.018	0.003
30	0.027	0.009	0.024	0.004
35	0.034	0.012	0.030	0.004
40	0.042	0.015	0.037	0.004
45	0.051	0.018	0.044	0.006
50	0.061	0.020	0.053	0.006
55	0.072	0.024	0.062	0.008
60	0.085	0.028	0.073	0.011
65	0.100	0.033	0.085	0.013
70	0.117	0.038	0.099	0.018
75	0.137	0.044	0.116	0.022
80	0.161	0.052	0.135	0.028
85	0.191	0.061	0.160	0.035
90	0.230	0.074	0.193	0.048
95	0.291	0.093	0.243	0.069
100	0.530	0.166	0.450	0.169

Table SI3.10: Fraction of 100 kb windows that contain $\frac{i}{20}, i \in \{1, \ldots, 20\}$ of the Neandertal ancestry as measured by $ta_{0.25}$.

SI 4 Tiling path of Neandertal haplotypes

One possibility offered by the map of Neandertal ancestry is that we can exploit modern human genomes to reconstruct the genome of the introgressing Neandertal. To do so, we used the inferred Neandertal haplotypes as defined in Section SI 3.1. We chose haplotypes that are at least 0.02 cM as inference of longer haplotypes are shown to be even more accurate in simulations (Section SI 2.3). At each SNP which is covered by at least one Neandertal haplotype, we reconstructed the Neandertal base as the consensus allelic state across all the inferred haplotypes (see Extended Data Fig. 3a).

Applying this procedure in each of the 1000 Genomes European and East Asian populations and merging the haplotypes, we reconstructed 4437 Neandertal contigs that cover a total length of 1.1 Gb. The median length of the contigs is 129 Kb (see Extended Data Fig. 3b). To convert these numbers into percentages, we determined the number of bases in the autosomes of the human genome reference (GRCh37) (2.68 Gb). Combining this tiling path with the high coverage Neandertal genome (Prüfer et al., 2013) infers a total of 89.8% or 2.41 Gb of the euchromatic Neandertal genome

SI 5 Tests for Positive Selection on Neandertal variants

In this section, we formulate a procedure to scan for Neandertal variants that may have been positively selected. The basic idea behind this scan is to estimate the distribution of the frequency of Neandertal variants under a model of neutral drift. Given this distribution, we can scan for regions of the genome at which the observed frequency of Neandertal ancestry is higher than expected under neutrality. Our procedure has two components:

- Estimate the background distribution of neutral Neandertal alleles introgressed into a modern human population.
- Use this distribution to estimate the tail probability that the frequency of Neandertal ancestry in a region exceeds the observed Neandertal frequency under neutrality. The Neandertal frequency in a region is obtained from the individual-level Neandertal ancestries that are estimated using the CRF.

SI 5.1 Summary

- To estimate the background distribution, we first estimate the Neandertal frequency spectrum *i.e.*, the frequency spectrum of introgressed Neandertal alleles. We fit a simple model that relates the Neandertal frequency spectrum to the initial frequency, α, and the drift τ experienced by introgressing Neandertal alleles. In this model, Neandertal alleles entered the modern human population at a frequency equal to α and then drifted neutrally till the present. We fit this model to a subset of the spectrum that is informative of drift since gene flow (we term this the SUBSET-estimator). The reason for looking at a subset of the spectrum is that our procedure for estimating the Neandertal frequency spectrum also includes a contribution from non-Neandertal alleles (we explain this in greater detail in Section SI 5.2.1). By restricting our analysis to a subset of the spectrum, we obtain more accurate estimates of drift.
- On data simulated under a range of demographic models, this estimator provides an adequate null model in a scan for positive selection *i.e.*, the tail probabilities estimated using the model exceed the tail probabilities observed in simulated data. Thus, P-values estimated using this model will be conservative.

  We test the estimator under demographic models that include models of constant population size, bottlenecks, recent population expansion as well as non-random mating of the non-African population (this includes recent admixture from Africans, and additional gene flow from Neandertals). Thus, although our model for estimating drift ignores the complexities of human history, it is robust to model misspecification.
- We applied this procedure to the 1000 Genomes Project data. We find that the maximum likelihood estimates of *drift* are consistent across subpopulations. For example, we estimate apparent drifts of $0.07 - 0.08$ in Europeans. These estimates are consistent whether we analyze all European individuals in 1000 Genomes Project or we restrict to the CEU population. They are also consistent when we analyze the Neandertal frequency spectrum estimated off genotypes or directly from the sequencing reads. Similarly, we obtain estimates of drift in East Asians of about 0.10 (when we analyze all East Asian individuals and only the CHB). Our analysis of the American populations does not reveal substantially more drift than Europeans – we estimate drift of about 0.08 when we analyze all Americans or Mexicans alone (see Table SI 5.2). This result could arise because the European component of the ancestry of American populations is dominating the signal.

- Using the maximum likelihood estimates $(\hat{\alpha}, \hat{\tau})$, we scanned the European and East Asian populations in non-overlapping 100 Kb windows. We consider the average Neandertal ancestry la in 100 Kb non-overlapping windows restricting to windows that contain at least 10 SNPs that pass filters. Within each window w, we estimated the average Neandertal ancestry

  $$la(w) = \frac{\sum_{j \in S(w)} \sum_{s=1}^{m} \gamma_{s,j}}{m|\{j \in S(w)\}|}$$

  Here m, refers to the number of haplotypes, $\gamma_{s,j}$ refers to the marginal probability estimated by the CRF at SNP j in haplotype s, and $S(w)$ refers to the set of SNPs that belong to window w. Within each window, we estimated the P-value for the Neandertal ancestry drifting to a more extreme value than the observed average Neandertal ancestry la in that window.

  Applying this procedure to the European individuals in 1000 Genomes, we identified 10 regions that are significant at FDR<0.10. Of these, 4 regions passed the Bonferroni corrected P-value threshold of 0.05 .

  In the combined East Asian data, there are 12 regions that are significant at FDR $<$ 0.10. 3 regions passed the Bonferroni significance threshold.

- Our analysis in Section SI 8 indicates that the assumption of neutrality is inappropriate. To test for the robustness of our estimator to this violation, we divided the genome into quintiles based on the B-statistic and estimated the background distribution in each quintile. We find in practice that the bin of highest B-statistic has both the highest mean Neandertal ancestry as well as the highest variance in Neandertal ancestry across loci. Therefore, we considered two approaches to deal with the effects of non-neutrality. In one approach, we assume that the quintile with the highest B-statistic has not been affected by purifying selection. Under this assumption, we can use the background distribution estimated from this quintile in our scan and be confident that the statistics are conservative. In the second approach, we assign a P-value to a region based on the background distribution of regions of similar B-statistic.

  The first approach yields 4 regions as significant in EUR at $FDR < 0.1$ (2 of which passed the Bonferroni significance threshold) while no regions are significant in ASN. Using the second approach, we see 20 and 24 regions that are significant at $FDR < 0.1$ in EUR and ASN respectively with 4 regions passing the Bonferroni significance threshold in each population.

SI 5.2 Estimating the Neandertal frequency spectrum

To devise a test for positive selection, we estimate the Neandertal frequency spectrum *i.e*, the frequency spectrum of Neandertal alleles that have introgressed. One strategy to estimate the Neandertal frequency spectrum would be to use the estimates of Neandertal ancestry from the CRF. However, a potential pitfall with this approach is that false positives and false negatives associated with the CRF might bias our estimate of the frequency spectrum. Instead, we estimate the frequency spectrum of Neandertal alleles directly from related site frequency spectra.

To estimate the Neandertal frequency spectrum, we make use of the recently sequenced high-coverage Denisova genome (Meyer et al., 2012). Neandertals and Denisovans are approximate sister groups (Reich et al., 2010; Meyer et al., 2012) although there is evidence that Denisovans have a small fraction of their ancestry from an archaic population not related to Neandertals (Prüfer et al., 2013). Unlike the Neandertals, Denisovans have been shown to have not contributed substantial ancestry to European and East Asian populations although small levels of gene flow have been reported (Skoglund and Jakobsson, 2011; Prüfer et al., 2013). We use this differential relationship of Neandertals and Denisovans to estimate the Neandertal frequency spectrum. We consider two spectra – the site frequency spectrum in a non-African population conditioned on ascertaining a derived allele in Neandertal and an ancestral allele in Denisova, $nd10$, and the site frequency spectrum conditioned on ascertaining a derived allele in Denisova and an ancestral allele in Neandertal, $nd01$.

We computed the $nd10$ and $nd01$ spectra as well as their difference δ in the 1000 Genomes Project populations. Figures SI 5.1 and SI 5.2 show these spectra. We observe that δ is qualitatively different in the African vs the non-African populations. We use δ as an estimate of the spectrum of introgressed Neandertal alleles. To understand why this is a valid interpretation, consider the class of SNPs that are not introgressed. These SNPs were either polymorphic in the modern human lineage at the time of introgression or were fixed derived – otherwise we would not observe the derived alleles in the archaic genomes. In the latter case, the effect of gene flow is found in the high-frequency end of the spectrum which we ignore. In the former case, the SNP must have been polymorphic in the modern human-Neandertal ancestor and would have undergone significant drift on both the modern human and the archaic lineages. As a result, the frequency of such a SNP in non-Africans would approximately be independent of the state on the archaic ($nd10$ vs $nd01$) and would not contribute to the difference spectrum δ.

SI 5.2.1 Estimating the neutral model

Our data consists of the difference spectrum in a non-African population in which we randomly sampled n chromosomes, $\delta(i) = nd10(i) - nd01(i)$, where $i \in \{1, \ldots, n-1\}$ denotes the number of copies of the derived allele observed. Our procedure to estimate drift since Neandertal gene flow makes the following simplifying assumptions:

- The majority of the Neandertal alleles are evolving neutrally.
- A Neandertal allele enters the ancestral European population at a frequency α. Thus, we ignore the distribution of allele frequencies at the time of introgression (although the allele frequency is upper bounded by the Neandertal admixture proportion). Following introgression, these alleles drift neutrally until the present. The dynamics of this drift are determined by a single drift parameter which encompasses the effects of the time since gene flow and the population sizes during this time $\tau = \int_0^t \frac{dt}{2N(t)}$. In effect, we ignore the effect of non-random mating, *e.g.*, due to later admixture events that have been shown to have occurred (Meyer et al., 2012; Wall et al., 2013).

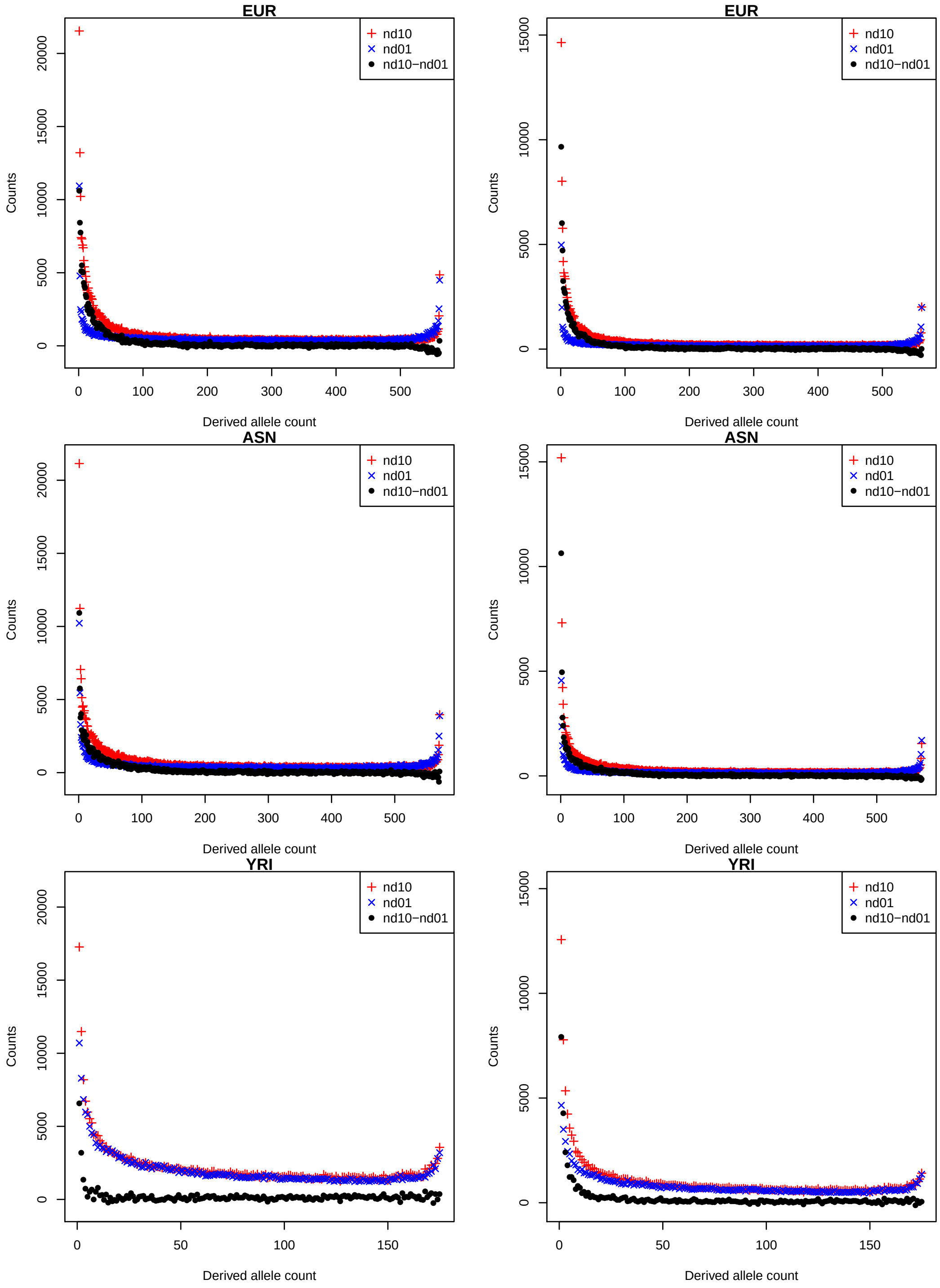
Assuming no linkage between sites, we can write (independently for each i) (Sawyer and Hartl, 1992)

$$\begin{aligned} nd10(i) &\sim \text{Pois}(\mu_{10}(i)), i \in \{0, \ldots, n-1\} \\ nd01(i) &\sim \text{Pois}(\mu_{01}(i)), i \in \{0, \ldots, n-1\}, \end{aligned}$$

The distribution of $\delta(i) = nd10(i) - nd01(i)$ is given by

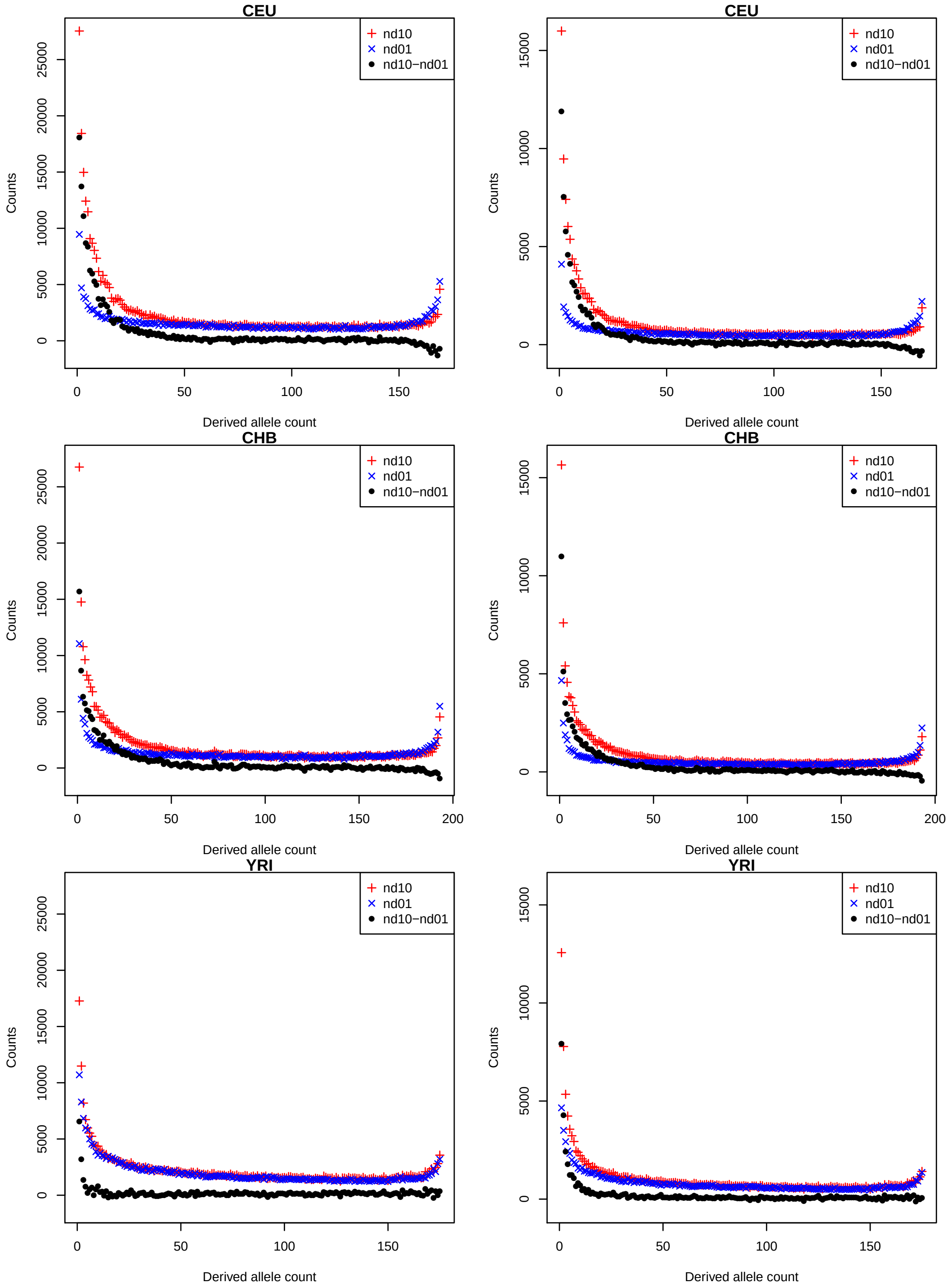
$$\begin{aligned} \delta(i) &\sim Skellam(\mu_{10}(i), \mu_{01}(i)) \\ &\sim \mathcal{N}(\Delta(i), S(i)) \end{aligned}$$

Here *Skellam* denotes the Skellam distribution which arises as the difference of two independent Poisson random variables. We use the normal approximation to the Skellam distribution which holds when the observed counts are large, as is the case here (Abramowitz and Stegun). Denote



(a) (b)

Figure SI 5.1: Conditional allele frequency spectra $nd10, nd01$ and an estimate of the Neandertal allele frequency spectrum, δ, in EUR, ASN and YRI computed using a) the high-coverage Neandertal and Denisova genomes and b) the low-coverage Neandertal and Denisova genomes.



(a) (b)

Figure SI 5.2: Conditional allele frequency spectra $nd10, nd01$ and an estimate of the Neandertal allele frequency spectrum, δ, in CEU, CHB and YRI computed using a) the high-coverage Neandertal and Denisova genomes and b) the low-coverage Neandertal and Denisova genomes.

$\Delta(i) = \mu_{10}(i) - \mu_{01}(i), S(i) = \mu_{10}(i) + \mu_{01}(i)$ where $\mu_{10}(i), \mu_{01}(i)$ are the expected number of sites in the non-African population with derived allele count i under the nd10 and nd01 ascertainments respectively. The quantity $S(i)$ is a function of the entire demographic history of modern non-Africans, Neandertals, and Denisovans. If this history were known, we could analytically compute $S(i)$. Since a detailed model of history is not known, we instead replace $S(i)$ by its plug-in estimator $\hat{S}(i) = nd10(i) + nd01(i)$.

$$\delta(i) \quad \sim \quad \mathcal{N}(\Delta(i), \hat{S}(i)) \tag{5}$$

$\Delta(i)$ is the quantity of interest – it is closely related to the expected counts of the introgressed Neandertal alleles. $\Delta(i)$ does not represent exactly the counts of Neandertal alleles. For example, a model in which Neandertals and Denisovans form a sister group (ignoring any recent gene flow) does not fit the data and analyses: (Prüfer et al., 2013) show that this observation is consistent with the Denisovan genome sequence being an admixture of a group related to Neandertals with an archaic population that split off from the ancestors of modern humans and Neandertals before the two groups diverged. Thus, in addition to the introgressed Neandertal alleles, $\Delta(i)$ includes a contribution from alleles that were segregating in the Neandertal-modern human ancestor, *e.g.*, mutations that arose after the split of the archaic population that contributed genes to the Denisovan genome but before the split of Neandertals and modern humans. We propose an estimator to deal with the noise in $\Delta(i)$.

SUBSET-estimator: In this estimation approach, we restrict our attention to bins of δ where the non-Neandertal component is unlikely to make a large contribution. We restrict to derived allele counts such that the derived allele frequency $\in [0.01, 0.10]$. This range of frequencies excludes singletons and often doubletons – this class of SNPs has been shown to have a non-negligible error rate and the AFS in these frequency bins might not be reliably estimated. This range also excludes higher frequency variants where the non-Neandertal component makes a dominant contribution. We have shown previously that this range of frequencies is enriched for Neandertal introgressed alleles (Sankararaman et al., 2012).

We can then model $\Delta(i), i = \{1, \ldots, n-1\}$ as

$$\Delta(i) \quad = \quad c\binom{n}{i}\int_0^1 dx x^i (1-x)^{(n-i)} K(x; \alpha, \tau) \tag{6}$$

Here $K(x; y, \tau)$ is the transition density function for the neutral Wright-Fisher diffusion with no mutations and denotes the transition density of the frequency y of an allele that starts at frequency y and drifts for τ units. $K(.;.,.)$ can be computed analytically (Kimura, 1955). c is a scaling factor. From Equations 5 and 6, we can estimate the parameters (c, α, τ) by maximizing the likelihood (see Appendix A) where the log likelihood function has the form

$$\mathcal{L}(c, \alpha, \tau) = -\sum_{\{i: 0.01 < \frac{i}{n} <= 0.1\}} \frac{(\delta(i) - \Delta(i; c, \alpha, \tau))^2}{2\hat{S}(i)} \tag{7}$$

SI 5.2.2 Neutral simulations to assess the adequacy of our procedure

We simulated 10 Gb worth of sequence data (10000 1Mb loci) under several demographic models that relate Europeans, Africans, Neandertal and Denisova using ms (Hudson, 2002). In these models, the Denisova sequence was modeled as an admixture of a population related to Neandertals and an archaic population. The archaic population split from the Neandertal-modern human ancestor 13000 generations ago and contributed 25% of the ancestry of the Denisova genome while the ancestors of

Neandertals and modern humans split 12000 generations ago. These parameters induce a drift of 0.05 on the population ancestral to Neandertals and modern humans since the split of the archaic population. We simulated gene flow from Neandertal into the ancestors of Europeans. The time of gene flow was set to 2000 generations ago. For each simulation, we estimated $(\hat{\alpha}, \hat{\tau})$ as described in Section SI 5.2.1. We also computed the true average Neandertal ancestry in non-overlapping 100 Kb windows and then estimated a nominal P-value or a tail probability for each window w, $\frac{\int_{la(w)}^{1} dy K(y;\hat{\alpha},\hat{\tau})}{\int_{0}^{1} dy K(y;\hat{\alpha},\hat{\tau})}$. We compared the theoretical tail probabilities to the observed tail probabilities restricting to windows for which the average Neandertal ancestry lies in $(0, 1)$. A conservative procedure should yield theoretical tail probabilities that are not smaller than the observed tail probabilities.

We considered several classes of demographic models:

- Simple: In these models, after a single gene flow event, the ancestral European population is randomly mating till the present. We consider variants of these models : constant effective population of I) 20000, II) 10000, III) 5000, IV) a model of constant effective population size 10000 with a bottleneck of duration 20 generations in which the effective population size was reduced to 100 and V) a model of constant effective population 10000 followed by exponential growth starting 400 generations in the past so that the current population size is one million. For each of these models, we considered admixture proportions of $\{0.01, 0.02, 0.03, 0.04\}$.

- Dilution: In this model, the European population has a constant effective population size of 10000. The European population experiences gene flow, 1500 generations ago, from an African population that reduces the mean Neandertal ancestry. We assumed that the proportion of dilution is 0.5. We varied the admixture proportion so that the Neandertal ancestry in present-day Europeans was $\{0.01, 0.02, 0.03, 0.04\}$.

- Double: In this model, the European population experiences two discrete gene flow events from the Neandertal at 2000 and 1500 generations respectively. The second gene flow event produces a date that is at the lower-end of the time of last exchange of genes estimated previously (Sankararaman et al., 2012). This model has two parameters : the proportion of Neandertal ancestry in present-day Europeans f and the proportion of Neandertal ancestry in present-day Europeans from the older gene flow which we set to 0.5. We vary f across $\{0.01, 0.02, 0.03, 0.04\}$.

Table SI 5.1 shows the calibration of the probabilities estimated using the estimator detailed in Section SI 5.2.1. For each simulated dataset, we used the difference spectrum δ to obtain maximum likelihood estimates of $(\hat{\alpha}, \hat{\tau})$. We then used these estimates to construct the maximum likelihood estimate of the Neandertal frequency spectrum $\hat{K} = K(.; (\hat{\alpha}, \hat{\tau}))$. We used the estimated spectrum, $\hat{K}$, to assign tail probability to the frequency of Neandertal ancestry over each non-overlapping 100 Kb window (for all windows with frequency of Neandertal ancestry in $(0, 1)$). For each $t \in \{0.05, 0.01, 10^{-3}, 10^{-4}\}$, we then assessed $o_{(1-t)}$, the fraction of windows with the frequency of Neandertal ancestry in $(0, 1)$ at which the tail probability is less than t. Table SI 5.1 lists the ratio $r_{(1-t)} = \frac{o_{(1-t)}}{t}$. A ratio ≤ 1 implies that the procedure to estimate the neutral frequency spectrum $\hat{K}$ is conservative and is an appropriate procedure to assess whether the Neandertal allele in a region is at an unexpectedly high frequency.

Table SI5.1 shows that the estimated frequency spectrum $\hat{K}$ is quite conservative across the range of models and parameter values considered. It is reassuring that even though our model makes a number of simplifying assumptions to estimate the neutral Neandertal frequency spectrum, the estimated spectrum is conservative in estimating the tail probability of a region attaining frequency above a threshold. Figures SI 5.3 and SI 5.4 depict the calibration curves over the entire range of the tail probabilities.

SI 5.2.3 Estimate of drift in 1000 Genomes data

We estimated the Neandertal frequency spectrum in several non-African populations in the 1000 Genomes Phase I project. We estimated the spectrum from the called genotypes. We randomly sample single alleles from each of the Neandertal and Denisova sequences to estimate the $nd10$ and $nd01$ spectra. For each of the populations analyzed, we estimated the maximum likelihood values of $(\hat{\alpha}, \hat{\tau})$ using the SUBSET estimator respectively.

While estimating the MLE of (α, τ) for the European and East Asian populations, we constrain α to lie within $[0.00, 0.04]$ since α is upper bounded by the Neandertal admixture proportion.

We estimated drifts using genotype data in several populations.

- 85 individuals from the CEU population
- 379 individuals from the EUR population
- 97 individuals from the CHB population
- 286 individuals from the ASN population
- 101 individuals from the MXL population
- 274 individuals from the AMR population

Table SI 5.2 shows the maximum likelihood estimates for each of these populations. We make several observations.

- The MLE is consistent across populations (*i.e.*, it is consistent whether we use CEU or EUR). We notice that α is the same value in all analyses. This is an artifact of the procedure used for maximizing the likelihood which uses an initial grid over which to maximize parameters followed by a subsequent refinement. Maximizing the likelihood directly using a Nelder-Mead Simplex algorithm yields only a slight improvement to the likelihood while the results of the subsequent selection scan are unchanged.

- The estimated drift in East Asians is larger than in Europeans. This is consistent with previous studies that have reported larger drift in East Asians than in Europeans since they diverged from each other (Gutenkunst et al., 2009; Keinan et al., 2007). The estimated drift in the American populations is slightly larger than in Europeans. The interpretation of drift in a recently admixed population is not totally clear. For example, Mexicans consist of genetic contributions from European, Native American and African populations Bryc et al. (2010). The Native American populations have experienced large amounts of drift since their divergence from other European populations (Reich et al., 2012). On the other hand, the African components dilute the proportion of Neandertal ancestry and reduces the apparent genetic drift.

- One of the worries about these estimates of drift is that the spectra used are based on genotypes estimated from low-coverage sequencing data. A number of studies have shown that the direct use of genotypes from low or medium coverage sequencing data leads to a substantial bias in the estimate of the SFS (Nielsen et al., 2011; Li, 2011). To assess the robustness of our

Model	Admixture	Normalized tail probabilities			
	fraction	0.05	0.01	10^{-3}	10^{-4}
Simple I	0.01	0.28	0.17	0.21	0.21
	0.02	0.49	0.38	0.36	0.14
	0.03	0.72	0.54	0.28	0.00
	0.04	0.95	0.73	0.50	0.43
Simple II	0.01	0.24	0.16	0.03	0.00
	0.02	0.29	0.17	0.06	0.00
	0.03	0.44	0.33	0.14	0.00
	0.04	0.58	0.46	0.31	0.26
Simple III	0.01	0.07	0.02	0.00	0.00
	0.02	0.14	0.05	0.00	0.00
	0.03	0.15	0.08	0.02	0.00
	0.04	0.41	0.28	0.22	0.00
Simple IV	0.01	0.07	0.03	0.00	0.00
	0.02	0.24	0.19	0.09	0.30
	0.03	0.17	0.04	0.00	0.00
	0.04	0.24	0.13	0.02	0.00
Simple V	0.01	0.24	0.17	0.10	0.00
	0.02	0.33	0.20	0.08	0.19
	0.03	0.44	0.28	0.13	0.15
	0.04	0.60	0.42	0.16	0.13
Dilution I	0.01	0.20	0.12	0.08	0.00
	0.02	0.32	0.20	0.09	0.00
	0.03	0.37	0.22	0.07	0.00
	0.04	0.49	0.31	0.19	0.24
Double I	0.01	0.34	0.27	0.22	0.31
	0.02	0.45	0.38	0.24	0.19
	0.03	0.60	0.47	0.39	0.29
	0.04	0.75	0.62	0.50	0.38

Table SI5.1: Calibration of the tail probabilities of the frequency of Neandertal ancestry as estimated by the SUBSET estimator described in Section SI5.2.1. The estimated tail probabilities are compared to the tail probabilities observed under various demographic models. We used the procedure in Section SI 5.2.1 to assign tail probabilities to the frequency of Neandertal ancestry observed in non-overlapping 100 Kb windows. We computed the empirical tail probabilities at several cutoffs on the expected tail probabilities $(0.05, 0.01, 10^{-3}, 10^{-4})$. The empirical tail probability is estimated to be the fraction of 100 Kb windows (conditioned on windows with Neandertal ancestry frequency in $(0, 1)$) which are assigned estimated tail probabilities below the cutoff. We report the ratio of the empirical to the expected tail probabilities. A well-calibrated procedure should produce a ratio close to 1. A ratio less than 1 denotes that the procedure is conservative.

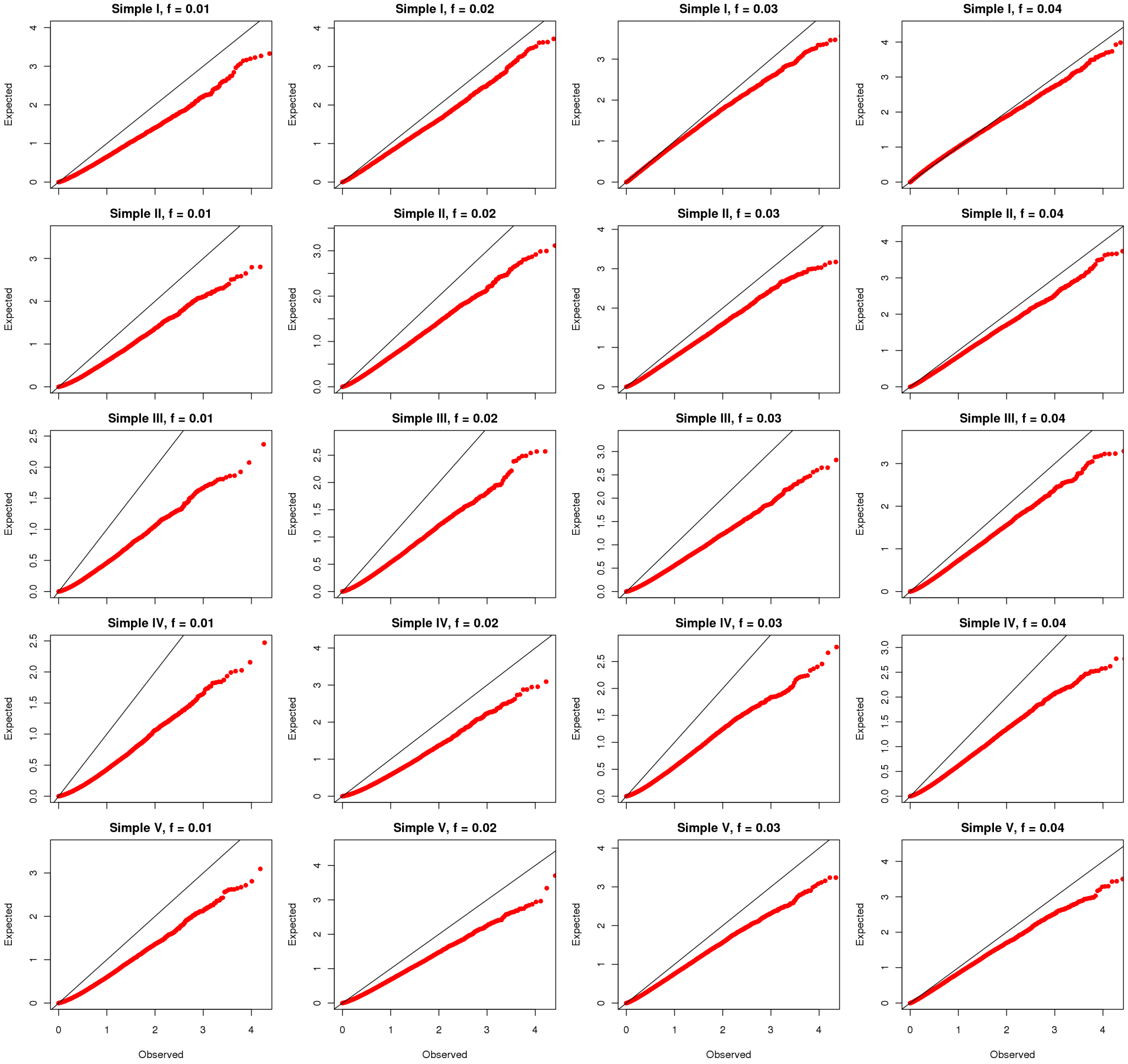


Figure SI5.3: Assessment of the calibration of the probabilities estimated using the SUBSET estimator from the model described in Section SI 5.2.1. The calibration was assessed using simulations. X and Y axes correspond to the observed and the expected tail probabilities, transformed by $-\log_{10}$. Each dot denotes the observed tail probability corresponding to an expected tail probability. A perfectly calibrated probability would lie along the diagonal line shown. Points that lie below the diagonal correspond to observed probabilities that are less than the expected probabilities from the model. Hence, these correspond to conservative estimates. We assessed calibration on the set of Simple models with varying admixture proportions $f \in \{0.01, 0.02, 0.03, 0.04\}$. We see that the model probabilities are conservative.

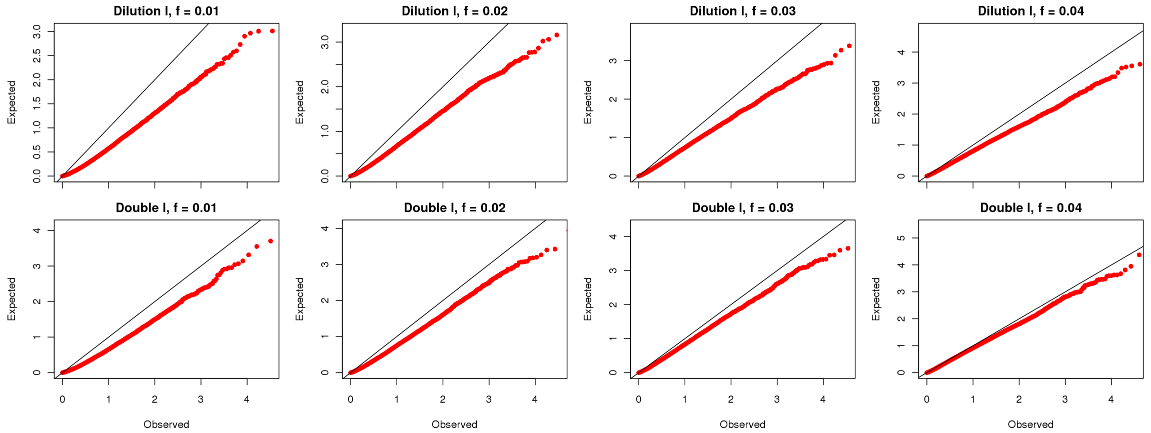


Figure SI 5.4: Assessment of the calibration of the probabilities estimated using the SUBSET estimator from the model described in Section SI 5.2.1. The calibration was assessed using simulations. X and Y axes correspond to the observed and the expected tail probabilities, transformed by $-\log_{10}$. Each dot denotes the observed tail probability corresponding to an expected tail probability. A perfectly calibrated probability would lie along the diagonal line shown. Points that lie below the diagonal correspond to observed probabilities that are less than the expected probabilities from the model. Hence, these correspond to conservative estimates. We assessed calibration on the set of Dilution and Double models with varying admixture proportions $f \in \{0.01, 0.02, 0.03, 0.04\}$. We see that the model probabilities are mostly conservative except for the Double I model with $f = 0.04$.

drift estimates to the procedure used to estimate the Neandertal spectrum, we re-estimated the SFS in the CEU population directly from sequence reads. To do so, we used samtools (Li et al., 2009) to compute genotype likelihoods. We then used bcftools to estimate the SFS using the EM-AFS method (Li, 2011). We ran 5 iterations of the EM. Table SI 5.2 shows that the estimates of drift when we estimate the SFS directly from the reads are very similar to the estimates obtained from genotypes. Thus, genotype calling on low-coverage data does not appear to seriously bias our estimates of drift.

SI 5.2.4 Selection scan

We use the maximum likelihood estimates $(\hat{\alpha}, \hat{\tau})$ to screen for positive selection on Neandertal alleles. To reduce the impact of false positives on this screen, we consider the average Neandertal ancestry la in 100 Kb non-overlapping windows restricting to windows that contain at least 10 SNPs that pass filters. As an estimate of the frequency of Neandertal ancestry, we chose to use the average the marginal probability, la, rather than the fraction of confidently called Neandertal alleles, *i.e.*, alleles with marginal probability above a threshold of 0.90 , $ta_{0.90}$. When we compared either of these statistics to the true frequency of Neandertal ancestry on simulated data (simulated under the model described in Section SI 2.1), we see that both statistics tend to underestimate the true Neandertal ancestry at regions with elevated Neandertal ancestry. The magnitude of underestimation is substantially more severe for the statistic that uses only the confidently called alleles, $ta_{0.90}$ (Figure SI 5.5).

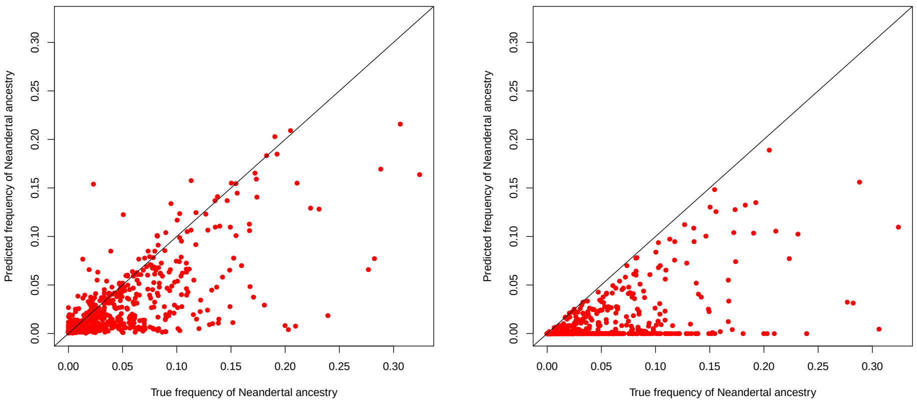
We then estimate a nominal P-value for each window w as the $\frac{\int_{la(w)}^{1} dy K(y;\hat{\alpha},\hat{\tau})}{\int_{0}^{1} dy K(y;\hat{\alpha},\hat{\tau})}$. Sometimes, for a given $(\hat{\alpha}, \hat{\tau}, la(w))$, the nominal P-value is estimated to be zero; in these cases, we set the nominal P-value to 10^{-9} because our algorithm to estimate K estimates K over a grid of size 1000 where the probability mass on each point in the grid is estimated to 6 digits of precision. We report windows that are significant at a False Discovery Rate < 0.10 using the Benjamini-Hochberg procedure to estimate the False Discovery Rate (Benjamini and Hochberg, 1995). We also report windows that attain a P-value less than 0.05 after Bonferroni correction (where the number of tests is taken to be the number of windows tested). We then merged consecutive windows that exceed the FDR significance threshold to identify putatively positively selected regions.

Applying this procedure to the European individuals in 1000 Genomes, we identified 10 regions that are significant at FDR of 0.10 using the SUBSET estimator. Table SI 5.3 list the regions that are significant at a $FDR < 0.1$. There are 4 that are significant at a Bonferroni-corrected P-value threshold of 0.05.

In the combined East Asian data, there are 12 that are significant at a FDR of 0.10. Table SI 5.4 list the regions that are significant at a $FDR < 0.1$ using the SUBSET. 3 regions passed the Bonferroni significance threshold.

We built a dendrogram summarizing the relationships of the haplotypes at one of the putatively positively selected regions - BNC2. See Figure SI 5.6.

To further analyze these regions, we chose the regions that were identified by the SUBSET estimator at a FDR < 0.10 in EUR and ASN *i.e.*, we chose 10 regions in EUR and 12 in ASN. Two of the regions (9:96,900,000-91,100,000 and 9:97,300,000-97,400,000) are identified in both Europeans and East Asians. We intersected these regions with regions identified using the CMS statistic (Grossman et al., 2010) on the 1000 Genomes Pilot data (cms). The CMS statistic combines several signals of selection to obtain greater power at detecting and localizing selective sweeps. We chose the regions that were identified by the CMS statistic in the CEU and the CHB+JPT populations. The regions were lifted over from hg18 to hg19 coordinates using the UCSC liftover tool (Hinrichs et al., 2006). We identified 2 regions that overlapped the CMS statistic. Both regions were found in EUR and had amongst the highest Neandertal frequencies (9:16.7-16.9 and 19:33.5-33.7). The non-overlap of all but two of the putatively positively selected loci with the CMS statistic suggest that using the variation in Neandertal ancestry could be sensitive to selective sweeps that are not easily detected by other statistics (Grossman et al., 2010). One possibility is that the distinctiveness of Neandertal ancestry allows us to detect signals of selection that are quite old (tens of thousands of years old) that are not easily detected by other signals. We also observe that none of these regions contain Neandertal alleles that have swept to fixation. Indeed, the frequencies of Neandertal ancestry in the putatively positively regions are in the range of $30-60\%$. This observation might be consistent with frequency-dependent selection on the introgressed Neandertal allele.



(a)

(b)

Figure SI 5.5: Comparison of estimates of the Neandertal ancestry to true Neandertal ancestry in simulations. We simulated 100 1Mb loci under the demographic model and parameters described in Section SI 2.1. We compared the estimates and true Neandertal ancestries. Panel (a) shows the estimates obtain from *la* while panel (b) shows the estimates from $ta_{0.90}$. $ta_{0.90}$ substantially underestimates the true Neandertal ancestry.

Data	Population	SUBSET estimator	
		$\hat{\alpha}$	$\hat{\tau}$
1000 Genomes genotypes	CEU	0.005	0.077
	EUR	0.005	0.073
	CHB	0.005	0.106
	ASN	0.005	0.106
	MXL	0.005	0.080
	AMR	0.005	0.083
1000 Genomes reads	CEU	0.0003	0.083

Table SI 5.2: Estimates of the model parameters (α,τ) on various populations sequenced in the 1000 Genomes Phase I data. We estimated these parameters using the SUBSET estimator. All, except one, of the estimates were computed from genotypes. To assess the effect of using genotypes, we also estimated these parameters using an AFS estimated directly from the reads for the CEU population.

Coordinates	Frequency		Pvalue	Genes
	la	$ta_{0.9}$	$-log_{10}(pval)$	
1:39.4-39.5	0.334	0.223	4.367	NDUFS5
1:57.2-57.3	0.303	0.180	3.892	C1orf168
1:170.3-170.4	0.309	0.297	3.983	
2:154.9-155.0	0.301	0.288	3.862	GALNT13
2:160.0-160.2	0.341	0.268	4.477	TANC1
2:238.8-239.0	0.457	0.418	**6.457**	SCLY
3:20.5-20.6	0.307	0.273	3.953	
4:28.3-28.4	0.326	0.308	4.243	
6:52.1-52.2	0.446	0.353	**6.252**	IL17F
6:66.4-66.7	0.368	0.355	4.909	
8:13.7-13.9	0.371	0.329	4.958	
8:14.0-14.2	0.350	0.290	4.619	SGCZ
9:16.7-16.9	0.637	0.545	**9.000**	BNC2
9:96.9-97.1	0.330	0.285	4.305	FAM22F
9:97.3-97.4	0.329	0.283	4.290	FBP2
12:52.9-53.0	0.367	0.250	4.892	KRT5,KRT71,KRT74,KRT72
12:113.4-113.5	0.314	0.201	4.059	OAS2
12:133.4-133.5	0.316	0.258	4.089	CHFR
14:44.8-44.9	0.305	0.266	3.922	
15:84.7-84.9	0.325	0.236	4.228	LOC100505679
15:85.8-86.1	0.398	0.367	5.406	AKAP13
16:78.0-78.1	0.354	0.330	4.683	CLEC3A
18:60.1-60.2	0.301	0.257	3.862	ZCCHC2
19:33.5-33.8	0.640	0.539	**9.000**	LRP3,SLC7A10,CEBPA,RHPN2

Table SI5.3: Regions of putative positive selection in Europeans when we use the SUBSET estimator. We list all regions that are significant with $FDR < 0.1$.

Coordinates	Frequency		Pvalue	Genes
	la	$ta_{0.9}$	$-log_{10}(pval)$	
1:208.5-208.6	0.417	0.298	3.894	
1:212.5-212.6	0.420	0.379	3.928	PPP2R5A
1:232.5-232.7	0.513	0.384	5.069	SIPA1L2
3:50.2-50.4	0.592	0.583	**6.175**	GNAT1,SLC38A3,GNAI2,SEMA3B C3orf45,IFRD2,NAT6,HYAL1 HYAL2,RASSF1,NPRL2,CYB561D2 TMEM115,TUSC2,ZMYND10,SEMA3F
3:191.2-191.3	0.433	0.348	4.080	
4:38.0-38.1	0.474	0.422	4.574	TBC1D1
4:38.2-38.8	0.514	0.418	5.082	TLR10
4:167.1-167.2	0.436	0.403	4.115	
4:167.3-167.4	0.418	0.410	3.905	
8:103.6-103.7	0.423	0.353	3.963	KLF10
9:90.7-90.8	0.419	0.291	3.917	FAM75C2
9:96.9-97.1	0.452	0.373	4.305	FAM22F
9:97.3-97.4	0.464	0.420	4.451	FBP2
9:112.8-113.2	0.571	0.456	**5.865**	AKAP2,C9orf152,TXN,TXNDC8
10:7.0-7.1	0.434	0.310	4.092	
11:99.1-99.2	0.479	0.453	4.635	
11:120.1-120.2	0.520	0.252	5.161	TMEM136
12:52.7-52.8	0.432	0.356	4.068	KRT86,KRT83,KRT85,KRT84
13:108.4-108.5	0.447	0.297	4.246	FAM155A
14:58.3-58.4	0.426	0.393	3.998	SLC35F4
20:62.1-62.4	0.620	0.492	**6.626**	PPDPF,PTK6,C20orf195,PRIC285 TNFRSF6B,ARFRP1,ZGPAT,LIME1 KCNQ2,SRMS,GMEB2,STMN3 RTEL1,RTEL1-TNFRSF6B,SLC2A4RG,EEF1A2
22:20.7-20.9	0.446	0.396	4.234	USP41,ZNF74,SCARF2,KLHL22

Table SI 5.4: Regions of putative positive selection in East Asians when we use the SUBSET estimator. We list all regions that are significant with $FDR < 0.1$.

SI 5.2.5 Robustness of signals of selection to assumption of neutrality

The procedure to scan for positive selection uses a model of drift that assumes that Neandertal introgressed variants are primarily neutral. The model assumes that these variants entered the modern human population at the same frequency and all experienced the same drift.

This model can be violated in several ways. Our analysis in Section SI8 indicates that Neandertal alleles near functionally important regions (as measured by a low value of the B-statistic (McVicker et al., 2009)) are likely to have been subject to strong purifying selection. Under this scenario, we might expect that the model of neutral evolution of Neandertal introgressed alleles would be particularly inappropriate for regions of low B-statistic where the effects of purifying selection are stronger; however, this model might still be appropriate in regions of high B-statistic where the introgressed Neandertal alleles are more likely to be evolving neutrally. One way to deal with this issue is to estimate the neutral model by restricting our analysis to the regions of the genome with high B-statistic.

B-statistic	EUR		ASN	
(quintile)	$\hat{\alpha}$	$\hat{\tau}$	$\hat{\alpha}$	$\hat{\tau}$
1	2.55e-04	0.072	3.09e-04	0.097
2	3.21e-04	0.075	1.00e-02	0.087
3	5.00e-03	0.065	5.00e-03	0.103
4	5.00e-03	0.075	4.29e-04	0.104
5	5.00e-03	0.084	2.49e-02	0.114

Table SI 5.5: Estimates of the model parameters (α,τ) on the EUR and ASN populations in quintiles of B-statistic (1-low,5-high). We estimated these parameters using the SUBSET estimator.

Coordinates	Frequency		Pvalue	Genes
	la	$ta_{0.9}$	$-log_{10}(pval)$	
2:238.9-239.0	0.457	0.418	5.337	SCLY
6:52.1-52.2	0.446	0.353	5.178	IL17F
9:16.7-16.8	0.637	0.545	**9.000**	BNC2
19:33.5-33.7	0.640	0.539	**9.000**	LRP3,RHPN2,GPATCH1,WDR88

Table SI 5.6: Regions of putative positive selection in Europeans when we fit the SUBSET estimator to the regions of the genome in the highest quintile of B-statistic. We list all regions that are significant with $FDR < 0.1$.

Alternately, even in the absence of purifying selection on Neandertal alleles, Neandertal alleles in a low B-statistic region are subject to larger drift due to the effects of background selection (Charlesworth et al., 1995). To deal with this issue, we could estimate a neutral model within bins of B-statistic. We then assign a P-value to a region using the model selected based on the B-statistic for the region.

To assess the impact of this issue, we partitioned the autosomes into quintiles of B-statistics. Each SNP in the 1000 Genomes dataset was assigned to one of the quintiles. We then applied the SUBSET estimator to estimate parameters for each of the B-statistic quintiles. Table SI 5.5 shows that estimates of the drift $\hat{\tau}$ decreases in regions of low B consistent with the action of purifying selection.

To deal with the effect of purifying selection, we used the estimates in the quintile of highest B-statistic to again scan for non-overlapping 100 Kb windows that are significant at FDR < 0.10. Using this estimator, 4 regions in EUR are significant at $FDR < 0.1$ while no regions are significant in ASN (Table SI 5.6).

As a second approach, we assigned a B-statistic to each non-overlapping 100 Kb window corresponding to the average B-statistic within the window. Each window was then assigned to a quintile based on the distribution of B-statistics. We then assigned a P-value to each window based on the estimates for the corresponding quintile (Table SI 5.5). Q-values and Bonferroni-corrected P-values were obtained from the P-values across all quintiles. Using this procedure, we see 20 and 24 regions that are significant at $FDR < 0.1$ in EUR and ASN respectively (Table SI 5.7 and SI 5.8).

These analyses show that an assessment of positive selection on Neandertal alleles is sensitive to assumptions about the neutral evolution of Neandertal alleles. Given the observation of purifying selection on Neandertal alleles in regions that are proximal to functional elements, it is clear that the background model of neutral evolution is inappropriate (at least in regions of the genome with

Coordinates	Frequency		Pvalue	Genes
	la	$ta_{0.9}$	$-log_{10}(pval)$	
1:39.4-39.5	0.334	0.223	4.316	NDUFS5
1:216.8-216.9	0.289	0.252	3.977	ESRRG
2:154.9-155.0	0.301	0.288	4.164	GALNT13
2:160.0-160.2	0.341	0.268	4.637	TANC1
2:238.8-239.0	0.457	0.418	**6.991**	SCLY
6:52.1-52.2	0.446	0.353	**9.000**	IL17F
6:66.4-66.5	0.368	0.355	4.567	
8:13.8-13.9	0.371	0.329	4.147	
9:16.7-16.9	0.637	0.545	**9.000**	BNC2
9:96.9-97.1	0.330	0.285	4.628	FAM22F
9:97.3-97.4	0.329	0.283	4.240	FBP2
12:52.9-53.0	0.367	0.250	4.829	KRT5,KRT71,KRT74,KRT72
12:113.4-113.5	0.314	0.201	4.370	OAS2
12:133.4-133.5	0.316	0.258	4.240	CHFR
14:44.8-44.9	0.305	0.266	4.228	
15:84.7-84.9	0.325	0.236	4.179	LOC100505679
15:85.8-86.1	0.398	0.367	5.384	AKAP13
16:78.0-78.1	0.354	0.330	5.022	CLEC3A
18:60.1-60.2	0.301	0.257	4.164	ZCCHC2
19:33.5-33.7	0.640	0.539	**9.000**	LRP3,RHPN2,GPATCH1,WDR88

Table SI 5.7: Regions of putative positive selection in Europeans when we fit the SUBSET estimator stratified by quintile of B-statistic. We list all regions that are significant with $FDR < 0.1$.

low B). As a result, the formal P-values assigned under such a model are unlikely to be meaningful.

Coordinates	Frequency		Pvalue	Genes
	la	$ta_{0.9}$	$-log_{10}(pval)$	
1:210.3-210.5	0.410	0.364	4.288	SYT14
1:212.4-212.6	0.420	0.379	4.416	PPP2R5A
1:232.5-232.7	0.513	0.384	5.029	SIPA1L2
2:68.3-68.4	0.371	0.255	3.804	WDR92
3:50.2-50.4	0.592	0.583	**9.000**	GNAT1,SLC38A3,GNAI2,SEMA3B C3orf45,IFRD2,NAT6,HYAL1 HYAL2,RASSF1,NPRL2,CYB561D2 TMEM115,TUSC2,ZMYND10,SEMA3F
4:38.0-38.1	0.474	0.422	4.545	TBC1D1
4:38.3-38.9	0.514	0.418	5.372	TLR10,TLR1,FAM114A1,KLF3
5:148.8-148.9	0.375	0.228	3.852	CSNK1A1
8:103.6-103.7	0.423	0.353	3.950	KLF10
9:96.9-97.1	0.452	0.373	4.284	FAM22F
9:97.2-97.4	0.464	0.420	4.997	HIATL1
9:112.8-113.2	0.571	0.456	**6.443**	AKAP2,C9orf152,TXN,TXNDC8
9:129.5-129.6	0.371	0.228	3.804	ZBTB43
10:69.3-69.4	0.413	0.369	4.326	CTNNA3
10:69.5-69.6	0.410	0.335	4.231	DNAJC12
11:99.1-99.2	0.479	0.453	4.803	
11:120.1-120.2	0.520	0.252	**5.793**	TMEM136
12:52.7-52.8	0.432	0.356	4.517	KRT86,KRT83,KRT85,KRT84
14:58.3-58.4	0.426	0.393	3.984	SLC35F4
15:79.0-79.1	0.389	0.259	4.025	ADAMTS7
16:89.7-89.8	0.387	0.311	4.000	CHMP1A,C16orf55,CDK10,SPATA2L C16orf7,DPEP1,ZNF276,LOC100128881
18:55.1-55.2	0.400	0.318	4.106	ONECUT2
20:62.1-62.4	0.620	0.492	**9.000**	PPDPF,PTK6,C20orf195,PRIC285 TNFRSF6B,ARFRP1,ZGPAT,LIME1 KCNQ2,SRMS,GMEB2,STMN3 RTEL1,RTEL1-TNFRSF6B,SLC2A4RG,EEF1A2
22:20.7-20.9	0.446	0.396	4.611	USP41,ZNF74,SCARF2,KLHL22

Table SI5.8: Regions of putative positive selection in East Asians when we fit the SUBSET estimator stratified by quintile of B-statistic. We list all regions that are significant with $FDR < 0.1$.

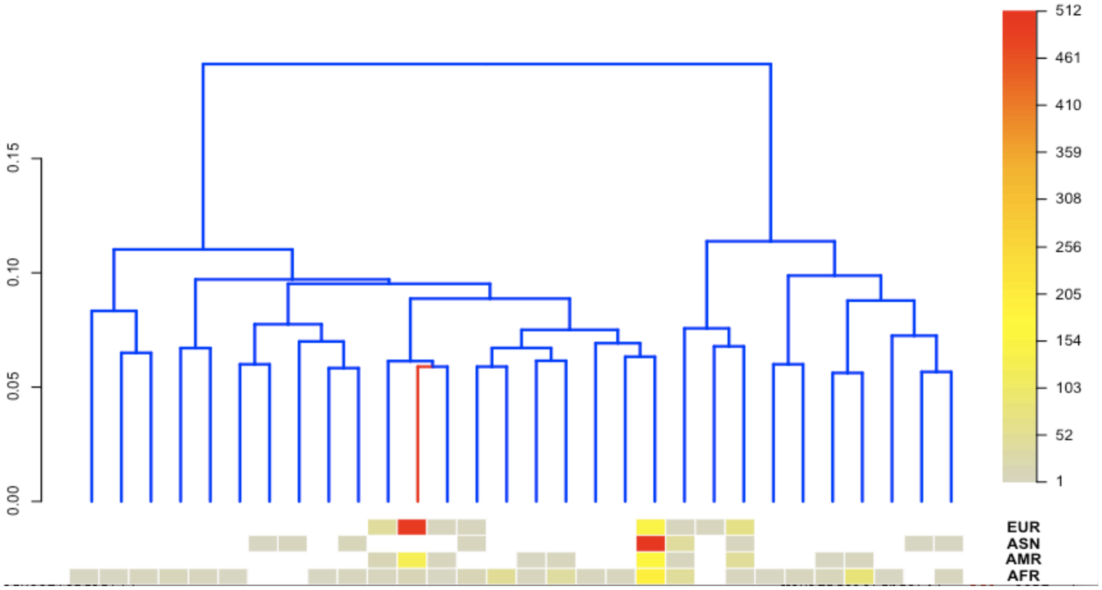


Figure SI 5.6: Dendrogram of 1000 Genomes haplotypes at the BNC2 region (chr9:16720122-16769662). We constructed a hierarchical average-linkage clustering of the haplotypes from the 1000 Genomes populations and archaic genomes (Altai and Vindija Neandertals and the Denisova genome). For the archaic genomes, we sampled a single allele at each SNP to form a haplotype. For the modern human genomes, we used computationally phased haplotypes. To aid visualization, we collapsed all internal nodes of the dendrogram at a height < 0.055 into a single leaf node. The red-colored edge of the dendrogram leads to the clade that contains both the Neandertal haplotypes. Below each leaf, we also list the distribution of all haplotypes across populations (EUR, ASN, AMR and AFR) that map to a clade. The Neandertal clade contains haplotypes that are present at high-frequency in EUR and ASN, is absent in ASN and is present in low-frequency in AFR.

SI 6 Functional implications of variants introgressed from Neandertals

A functional analysis of the genomic regions that are either significantly enriched for or significantly depleted of Neandertal ancestry in present-day human populations outside of Africa may provide some insights into phenotypes that have been important in recent human history.

SI 6.1 Correlation of Neandertal ancestry estimates across populations

We classified CCDS genes (Pruitt, Harrow et al. 2009) as of either high or low Neandertal ancestry based on the marginal probabilities of Neandertal ancestry inferred by the Conditional Random Field for each site within the gene. Specifically, the CRF estimates the probability of Neandertal ancestry at each SNP in an individual haplotype. We interpolate these probabilities to all bases in the genome using linear interpolation based on the physical distance.

We then define the two sets as follows:

- Genes with low Neandertal ancestry. A gene where all sites across all individuals are assigned a marginal probability of Neandertal ancestry <=10% is defined to have low Neandertal ancestry. Such genes are likely to be devoid of Neandertal ancestry in the sample analyzed. We use a cutoff of 10% to reduce false positives i.e., genes that have some Neandertal ancestry but are predicted to be low in Neandertal ancestry because the power to infer Neandertal ancestry is reduced in the region containing the gene. In addition, genes may falsely appear to be low in Neandertal ancestry due to low SNP density. To handle this, we exclude genes that contain fewer than 100 SNPs within a 100kb window centered at its midpoint.
- Genes with high Neandertal ancestry. To identify these genes, we computed an average introgression score for each of these genes as the average of the marginal probability across all individuals at all bases within the gene. Genes ranked in the top 5% of genes with evidence of Neandertal introgression were declared to have high Neandertal ancestry.

Using the averaged introgression scores for each CCDS gene, we tested the correlation of the values per gene between East Asians and Europeans. We found a strong correlation (ρ =0.71, p<2.2e-16), indicating that the frequency of introgressed alleles is similar between populations. In order to further investigate whether the high correlation between populations is driven by genes that are low or high in Neandertal ancestry, we assessed the overlap between the two populations for genes low in Neandertal ancestry. Testing 18017 genes, we find that 8475 genes have low Neandertal ancestry in Europeans and 9300 have low Neandertal ancestry in East Asians. A total of 6641 genes overlap between sets, significantly more than expected (p<2.2e-16, binomial test, two-sided). When considering genes with high Neandertal ancestry in Europeans (862 genes) and East Asians (862 genes), we find an equally strong overlap between populations (279 genes, p<2.2e-16, Fisher's exact test). We note that a significant correlation might be caused by the correlated variation in power in Europeans and East Asians. However, SI 10 shows that the correlation in Neandertal ancestry proportions increases at larger distance scales where power is expected to be more homogeneous, suggesting that power alone does not explain the correlation.

SI 6.2 Analysis of regions of low Neandertal ancestry

We assessed whether particular functional categories are over-represented among the genes determined to have either high or low Neandertal ancestry in present-day European and East Asian populations. For this we tested for enrichment in Gene Ontology (Ashburner, Ball et al. 2000) categories using the hypergeometric test implemented in the FUNC package (Prufer, Muetzel et al. 2007) and for enrichment in KEGG pathways (Kanehisa and Goto 2000) using the hypergeometric test implemented in GOStats (Falcon and Gentleman 2007) . For GO enrichment the FWER is calculated based on 1000 permutations and categories that are significant (FWER <0.05) in either population are reported. For the KEGG enrichment analysis no multiple testing correction has been applied. Categories reported are those with a raw p-value of <0.05 in either population.

For all functional enrichment tests, the correlation between the enrichment scores in categories found in Europeans and East Asians was calculated using Spearman correlation. For the regions that have low Neandertal ancestry, we find a consistently high and significant positive correlation between populations (ρ=0.44, p-value<2.2e-16 for GO and ρ=0.78, p-value<2.2e-16 for KEGG).

SI 6.2.1 Functional enrichment in regions of low Neandertal ancestry

The functional enrichment test identifies 33 Gene Ontology categories. These contain genes that carry out a number of basic cellular functions including genes involved in the cell cycle and RNA processing, as well genes that encode basic cellular structures such as the ribosomal proteins (Table SI 6.1). The 34 KEGG pathways that are identified (Table SI 6.2) indicate enrichment in a number of metabolic-related categories including starch and sucrose metabolism, specifically the amylases (AMY1, AMY2). It is known that amylase expansion has been important in modern human dietary adaptation. (Prüfer, Racimo et al. 2013) have shown that alpha-amylase has just 2 copies in Neandertals, though the family is expanded in modern humans. Olfactory transduction pathway genes are also highly enriched among the regions that are low in Neandertal ancestry.

GO_domain	GO_id	GO_term	EUROPEAN	EAST ASIAN
molecular_function	GO:0003676	nucleic acid binding	0.018	0.032
biological_process	GO:0006396	RNA processing	0.004	0.049
cellular_component	GO:0030529	ribonucleoprotein complex	<0.001	0.027
cellular_component	GO:0044422	organelle part	<0.001	0.037
cellular_component	GO:0044446	intracellular organelle part	<0.001	0.025
biological_process	GO:0016071	mRNA metabolic process	<0.001	0.014
cellular_component	GO:0031981	nuclear lumen	0.039	0.017
cellular_component	GO:0032991	macromolecular complex	<0.001	0.21
cellular_component	GO:0043226	Organelle	<0.001	1
cellular_component	GO:0043227	membrane-bounded organelle	0.015	1
cellular_component	GO:0043229	intracellular organelle	<0.001	1

cellular_component	GO:0043231	intracellular membrane-bounded organelle	0.041	1
cellular_component	GO:0044428	nuclear part	0.005	0.022
molecular_function	GO:0003723	RNA binding	0.036	0.26
molecular_function	GO:0004984	olfactory receptor activity	1	<0.001
biological_process	GO:0000956	nuclear-transcribed mRNA catabolic process	0.011	1
biological_process	GO:0006401	RNA catabolic process	0.021	1
biological_process	GO:0006402	mRNA catabolic process	0.021	1
biological_process	GO:0007049	Cell cycle	0.63	0.002
biological_process	GO:0022402	Cell cycle process	0.64	0.003
biological_process	GO:0022403	Cell cycle phase	0.98	0.015
biological_process	GO:0031424	keratinization	0.029	1
cellular_component	GO:0005622	intracellular	0.028	1
cellular_component	GO:0005634	nucleus	0.041	0.23
cellular_component	GO:0005740	mitochondrial envelope	0.049	0.056
cellular_component	GO:0005743	mitochondrial inner membrane	0.016	0.12
cellular_component	GO:0022626	cytosolic ribosome	0.041	1
cellular_component	GO:0031966	mitochondrial membrane	0.064	0.035
cellular_component	GO:0043228	Non-membrane-bounded organelle	0.005	0.97
cellular_component	GO:0043232	intracellular non-membrane-bounded organelle	0.005	0.97
cellular_component	GO:0044424	intracellular part	0.015	1
cellular_component	GO:0044429	mitochondrial part	0.006	0.77
cellular_component	GO:0044445	cytosolic part	0.003	0.99

Table SI 6.1: Gene Ontology functional enrichment categories for genes with low Neandertal ancestry. Columns EUR and ASN give the multiple testing corrected p-values in Europeans and East Asians, respectively.

KEGGID	Term	EUROPEAN	EAST ASIAN
3010	Ribosome	0.0014	0.031
4740	Olfactory transduction	0.008	4.40E-09
3008	Ribosome biogenesis in eukaryotes	0.017	0.019
650	Butanoate metabolism	0.018	0.034
3018	RNA degradation	0.023	0.0071
5012	Parkinson's disease	0.0077	0.066
4142	Lysosome	0.0091	0.15
3013	RNA transport	0.013	0.022
603	Glycosphingolipid biosynthesis - globo series	0.021	0.058
4350	TGF-beta signaling pathway	0.024	0.25
40	Pentose and glucuronate interconversions	0.044	0.055

640	Propanoate metabolism	0.061	0.009
270	Cysteine and methionine metabolism	0.14	0.00092
500	Starch and sucrose metabolism	0.15	0.037
140	Steroid hormone biosynthesis	0.23	0.038
72	Synthesis and degradation of ketone bodies	0.01	0.3
4621	NOD-like receptor signaling pathway	0.015	0.64
330	Arginine and proline metabolism	0.016	0.061
4110	Cell cycle	0.022	0.38
760	Nicotinate and nicotinamide metabolism	0.027	0.088
533	Glycosaminoglycan biosynthesis - keratan sulfate	0.041	0.41
61	Fatty acid biosynthesis	0.046	0.29
3015	mRNA surveillance pathway	0.048	0.34
1100	Metabolic pathways	0.11	0.0027
830	Retinol metabolism	0.19	0.0055
982	Drug metabolism - cytochrome P450	0.19	0.0023
310	Lysine degradation	0.23	0.038
350	Tyrosine metabolism	0.25	0.0027
3050	Proteasome	0.44	0.0036
4916	Melanogenesis	0.73	0.013
71	Fatty acid metabolism	0.77	0.011
4340	Hedgehog signaling pathway	0.8	0.0092

Table SI 6.2: KEGG functional enrichment categories for genes with low Neandertal ancestry. Columns EUR and ASN give the raw p-values in Europeans and East Asians, respectively.

SI 6.2.2 Enrichment for testis-expressed genes in regions with low Neandertal ancestry

A prediction of the Dobzhansky-Muller hypothesis is that incompatibilities that cause hybrid sterility will be enriched in testis expressed genes (Orr and Turelli 2001). We therefore tested whether there is evidence that regions with low Neandertal ancestry are enriched for genes expressed in testis.

We analyzed the expression of tissue-specific genes in regions of high and low Neandertal ancestry using the Illumina BodyMap 2.0 data,(Derrien et al Gen. Res. 2012) which provides expression information for 16 tissues including testis. We analysed whether testis-specific genes are more often in devoid regions than all other genes.

We developed a tissue-specificity metric that defines for each tissue those genes that are significantly more highly expressed in that tissue than they are in any of the other 15 BodyMap tissues using the DESeq package (Anders and Huber, 2010) and a p-value cut-off of 0.05.

When testing for expression enrichment in the regions devoid of Neandertal ancestry, a major concern is that genes with similar function and/or expression distribution are spatially clustered. Since regions

devoid of Neandertal ancestry tend to be large, an overlap between these regions and genes with a specific expression pattern can therefore arise by chance.
To correct for this we circularized the chromosomes and randomly rotated the gene annotations.
We test three different sets of chromosomes: the whole genome (autosomes+X chromosome), X chromosome only, and autosomes only. For each test set we generated 1000 random samples by choosing random rotations for each chromosome in a set. For the X chromosome (where there are fewer than 1000 genes) we used all possible rotations. We tested how often tissue-specific genes overlap regions devoid of Neandertal ancestry by chance by comparing Fisher-exact p-values for the random samples to the real data. The values reported in Table SI 6.3 give the fraction of random tests that have as low or lower p-values than the real data.

Testis is the only tissue for which there is a significant enrichment in devoid regions when considering genes genome-wide (Table SI 6.3). In Europeans 46.7% of non testis-specific genes are in devoid regions, which is significantly less (p=8.1e-5) than the 52.5% of testis-specific genes that are located in regions of low Neandertal ancestry. In East Asians 51.3% of non testis-specific genes are in devoid regions which is significantly less (p=0.001) than the 56.1% of testis-specific genes that are located in regions of low Neandertal ancestry. For the comparison of testis-specific to all other genes on the autosomes and X-chromosome separately we see the same trend, although some comparisons using only European or East Asian populations do not reach significance.
The same analysis for regions in the top 5% of Neandertal ancestry shows no enrichment for tissue-specific expression.

	European			East Asian		
Tissue	genome-wide	chrX	autosomes	genome-wide	chrX	autosomes
Adipose	0.9322	0.9989	0.8138	0.9919	1	0.9458
Adrenal	0.5011	NA	0.5011	0.4244	NA	0.4244
Blood	0.9934	0.9836	0.9876	0.941	0.7295	0.9356
Brain	1	0.9979	1	1	1	1
Breast	0.9825	0.6258	0.9871	0.9974	0.9409	0.9962
Colon	0.6365	0.7735	0.6266	0.9358	0.9748	0.8851
Heart	0.9925	0.7075	0.9927	0.8035	0.5702	0.8147
Kidney	0.9996	0.1541	0.9998	0.9958	0.0801	0.9983
Liver	0.9949	0.9919	0.9865	0.9974	0.8567	0.9972
Lung	0.9552	0.6421	0.9585	0.9894	0.871	0.9863
Lymph	0.8797	0.6159	0.8989	0.9881	0.5147	0.9913
Ovary	0.8396	0.9501	0.8105	0.6181	0.912	0.5777
Prostate	0.9953	0.7865	0.9957	0.9987	0.7292	0.9989
skeletal muscle	0.9478	0.703	0.9457	0.8312	0.1037	0.8832
Testes	**0.0095**	0.1277	**0.016**	**0.018**	**0.0389**	0.0549
Thyroid	0.8585	0.6167	0.8795	0.8726	0.936	0.8562

Table SI 6.3: Enrichment of tissue-specific genes in Neandertal devoid regions. 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. NA means that there were no tissue-specific genes for this tissue on the X-chromosome. Of the sixteen tissues tested, only testis-specific genes are significantly enriched in the Neandertal devoid regions.

SI 6.3 Regions of high Neandertal ancestry

For genes within the top 5% of introgression scores in either population, we find a slightly weaker correlation between populations (ρ=0.21,p-value<2.2e-16 and ρ=0.22,p-value=0.09 for GO and KEGG, respectively) compared to categories enriched in genes with low Neandertal ancestry. However, there is no substantial difference between populations for either categories of genes that are enriched or categories of genes that are depleted for introgression.

For the genes with the top 5% of introgression scores, we examined in further detail the categories that are shared by both Europeans and East Asians. As before, we reported multiple-testing-corrected p-values for the GO categories, and raw p-values for the KEGG pathways. There are two KEGG pathways: *long-term depression*, and *galactose metabolism*, (Table SI 6.5a) that are significantly enriched in both European and East Asian populations as well as in an introgression map using the combined populations. We note that the sets of genes responsible for the enrichment in Europeans and East Asians are largely non-overlapping which may suggest that different sets of genes were recruited independently for the same phenotype.

Only one GO cellular component (keratin filament) is enriched for Neandertal introgression in all populations suggesting that introgressed Neandertal alleles may have been used by modern humans to adapt their skin and hair morphology to non-African environments to which Neandertals were pre-adapted (Table SI 6.4a). Interestingly, among the genes driving the significance for this category we find four genes that have an unusually high frequency of introgression in East Asians (Supplementary material SI 6) (KRT83, KRT84, KRT85, KRT86) and three completely different genes with high frequency in Europeans (KRT5, KRT71, KRT74). The set of keratins that have risen to high frequency in East Asians are all members of the hair keratin group, while those at high frequency in Europeans are epithelial keratins or without annotated function.

We show correlation between Neandertal ancestry and the B-statistic, with Neandertal ancestry decreasing in regions of functional importance (SI 8 and SI 9). Functional categories may therefore also not be randomly distributed with regard to B-statistic. We therefore developed a method to control for the influence of the B-statistic on Neandertal ancestry, and then carried out the GO and KEGG enrichment analyses using these B-statistic corrected ancestries.

All 17, 249 CCDS genes (genes on the X chromosome and genes without an assigned B-statistic were removed) were assigned a B-statistic (averaged across all nucleotides) and a random number between 0 and 1. Genes were sorted by B-statistic and by random number and assigned to 20 equal-sized bins based on B-statistic. Within each bin genes were then resorted by Neandertal ancestry and then by the random number, thus assigning genes a bin-corrected percentile of Neandertal ancestry.

Genes with the top 5% of Neandertal ancestry were then used to carry out the enrichment tests. We also tested genes with the top 5% of Neandertal ancestry without carrying out the B-statistic correction.
We tested for enrichment in Gene Ontology (Ashburner, Ball et al. 2000) categories using the hypergeometric test implemented in the FUNC package (Prufer, Muetzel et al. 2007) and for enrichment in KEGG pathways (Kanehisa and Goto 2000) using the hypergeometric test implemented in GOStats (Falcon and Gentleman 2007) .

We find that the categories *galactose metabolism* and *keratin filament* are significant in East Asians, and the combined population maps while the categories *long-term depression* is significant only in the combined population set (Table SI 6.4b and Table SI 6.5b). Thus, systematic differences in purifying selection across different gene categories cannot explain these results..

We explored whether enriched categories are unusual in their recombination rate or B-statistic, which might produce an artifactual signal of enrichment. To do this, we

i. labeled all genes according to their average recombination rate determined from the African-American recombination map (Hinch et al. Nature. 2011)

ii. labeled genes according to their average B-statistic (which reflects the strength of background selection) (McVicker et al. PLoS Genetics. 2009)

We find that the Family-Wise Error Rate (FWER) in the GO enrichment analysis is only weakly correlated with average recombination rate (Spearman correlation coefficient in Europeans ρ=-0.007,p-value=3.1e-15 and in East Asians ρ=-0.005, p-value=3.2e-09) or B-statistic per category (in Europeans ρ=0.004, p-value=8.3e-09,and in East Asians ρ=0.002 p-value=2.5e-05).

Thus, the enrichment signal is unlikely to be driven by either unusual recombination rates or background selection.

Non B-corrected

GO_domain	***GO_term***	***GO_id***	***FWER(<0.05)***
EUROPEAN			
Cellular_component	**keratin filament**	GO:0045095	<0.001
biological_process	collagen catabolic process	GO:0030574	0.007
biological_process	regulation of cytokine production involved in inflammatory response	GO:1900015	0.035
EAST ASIAN			
biological_process	cellular response to zinc ion	GO:0071294	<0.001
Cellular_component	**keratin filament**	GO:0045095	<0.001
biological_process	cellular response to cadmium ion	GO:0071276	0.003
biological_process	cellular response to metal ion	GO:0071248	0.009
biological_process	negative regulation of biological process	GO:0048519	0.014
biological_process	transforming growth factor beta receptor signaling pathway	GO:0007179	0.017
molecular_function	cadmium ion binding	GO:0046870	0.027
biological_process	response to zinc ion	GO:0010043	0.035
biological_process	regulation of transforming growth factor beta receptor signaling pathway	GO:0017015	0.04
biological_process	negative regulation of growth	GO:0045926	0.04
COMBINED			
biological_process	negative regulation of biological process	GO:0048519	<0.001
biological_process	cellular response to zinc ion	GO:0071294	<0.001
Cellular_component	**keratin filament**	GO:0045095	<0.001
biological_process	cellular response to cadmium ion	GO:0071276	0.002
biological_process	negative regulation of cellular process	GO:0048523	0.03
biological_process	regulation of cytokine production involved in	GO:1900015	0.03

	inflammatory response		
biological_process	cellular response to metal ion	GO:0071248	0.038
biological_process	cellular response to inorganic substance	GO:0071241	0.046

Table SI 6.4a: Gene Ontology Functional Enrichment for the genes in the top 5% of Neandertal ancestry not corrected for B-statistic. Multiple testing corrected p-values are given for enrichment in Europeans, East Asians and a combined European and East Asian population. Categories identified as significantly enriched in all populations are in bold.

B-corrected

GO_domain	*GO_term*	*GO_id*	*FWER(<0.05)*
EUROPEAN			
Cellular_component	nucleosome	GO:0000786	0.001
biological_process	regulation of cytokine production involved in inflammatory response	GO:1900015	0.035
Cellular_component	protein-DNA complex	GO:0032993	0.048
EAST ASIAN			
biological_process	cellular response to zinc ion	GO:0071294	<0.001
Cellular_component	keratin filament	GO:0045095	<0.001
biological_process	cellular response to cadmium ion	GO:0071276	0.004
biological_process	cellular response to metal ion	GO:0071248	0.015
molecular_function	cadmium ion binding	GO:0046870	0.023
COMBINED			
biological_process	cellular response to zinc ion	GO:0071294	<0.001
Cellular_component	keratin filament	GO:0045095	<0.001
biological_process	cellular response to cadmium ion	GO:0071276	0.003
molecular_function	zinc ion binding	GO:0008270	0.034
biological_process	regulation of cytokine production involved in inflammatory response	GO:1900015	0.037

Table SI 6.4b: Gene Ontology Functional Enrichment for the genes in the top 5% of Neandertal ancestry corrected for B-statistic. Multiple testing corrected p-values are given for enrichment in Europeans, East Asians and a combined European and East Asian population. Categories identified as significantly enriched in all populations are in bold.

Non B-corrected

KEGGID	*Term*	*Raw P-value*
EUROPEAN		

524	Butirosin and neomycin biosynthesis	0.00068
51	Fructose and mannose metabolism	0.0021
4974	Protein digestion and absorption	0.012
4730	**Long-term depression**	**0.015**
52	**Galactose metabolism**	**0.022**
591	Linoleic acid metabolism	**0.025**
5323	Rheumatoid arthritis	0.028
130	Ubiquinone and other terpenoid-quinone biosynthesis	0.032
4270	Vascular smooth muscle contraction	0.036
3030	DNA replication	0.043
4720	Long-term potentiation	0.05
EAST ASIAN		
604	Glycosphingolipid biosynthesis - ganglio series	0.0027
4114	Oocyte meiosis	0.0033
4960	Aldosterone-regulated sodium reabsorption	0.0048
5410	Hypertrophic cardiomyopathy (HCM)	0.0054
4662	B cell receptor signaling pathway	0.0071
5414	Dilated cardiomyopathy	0.008
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.0084
4140	Regulation of autophagy	0.0093
5214	Glioma	0.014
4622	RIG-I-like receptor signaling pathway	0.02
52	**Galactose metabolism**	**0.021**
4260	Cardiac muscle contraction	0.023
4660	T cell receptor signaling pathway	0.023
4110	Cell cycle	0.025
450	Selenocompound metabolism	0.026
4975	Fat digestion and absorption	0.028
4664	Fc epsilon RI signaling pathway	0.03
592	alpha-Linolenic acid metabolism	**0.031**
4974	Protein digestion and absorption	0.032
4972	Pancreatic secretion	0.037
5160	Hepatitis C	0.038
4630	Jak-STAT signaling pathway	0.038
565	Ether lipid metabolism	0.041
4912	GnRH signaling pathway	0.042
5222	Small cell lung cancer	0.044
4730	**Long-term depression**	**0.044**
4010	MAPK signaling pathway	0.045
COMBINED		
4730	**Long-term depression**	**2.00E-04**

524	Butirosin and neomycin biosynthesis	0.00064
4912	GnRH signaling pathway	0.0016
590	Arachidonic acid metabolism	0.0016
52	**Galactose metabolism**	**0.0036**
591	Linoleic acid metabolism	**0.0042**
592	alpha-Linolenic acid metabolism	**0.0043**
4270	Vascular smooth muscle contraction	0.0044
4960	Aldosterone-regulated sodium reabsorption	0.0046
4662	B cell receptor signaling pathway	0.0067
4722	Neurotrophin signaling pathway	0.008
565	Ether lipid metabolism	0.0089
4664	Fc epsilon RI signaling pathway	0.0094
51	Fructose and mannose metabolism	0.01
4010	MAPK signaling pathway	0.011
4972	Pancreatic secretion	0.013
5214	Glioma	0.014
4720	Long-term potentiation	0.015
604	Glycosphingolipid biosynthesis - ganglio series	0.021
4660	T cell receptor signaling pathway	0.022
4370	VEGF signaling pathway	0.023
4114	Oocyte meiosis	0.024
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.025
4971	Gastric acid secretion	0.027
4975	Fat digestion and absorption	0.027
564	Glycerophospholipid metabolism	0.029
4910	Insulin signaling pathway	0.036
4914	Progesterone-mediated oocyte maturation	0.042
5410	Hypertrophic cardiomyopathy (HCM)	0.044
4115	p53 signaling pathway	0.049

Table SI 6.5a: KEGG pathway enrichment categories for the genes in the top 5% of Neandertal ancestry not corrected for B-statistic. The raw p-values are given for Europeans, East Asians and a combination of Europeans and East Asians. Categories identified as significant in all populations are in bold.

B-corrected		
KEGGID	***Term***	***Raw P-value***
EUROPEAN		
5322	Systemic lupus erythematosus	2.90E-20
4974	Protein digestion and absorption	0.013
524	Butirosin and neomycin biosynthesis	0.024
51	Fructose and mannose metabolism	0.025
10	Glycolysis / Gluconeogenesis	0.03
130	Ubiquinone and other terpenoid-quinone biosynthesis	0.047
EAST ASIAN		
604	Glycosphingolipid biosynthesis - ganglio series	0.0035
450	Selenocompound metabolism	0.0045
4960	Aldosterone-regulated sodium reabsorption	0.0069
5160	Hepatitis C	0.011
4630	Jak-STAT signaling pathway	0.012
4660	T cell receptor signaling pathway	0.014
4664	Fc epsilon RI signaling pathway	0.015
4114	Oocyte meiosis	0.015
4150	mTOR signaling pathway	0.017
52	Galactose metabolism	0.026
4622	RIG-I-like receptor signaling pathway	0.028
5215	Prostate cancer	0.028
4662	B cell receptor signaling pathway	0.032
4975	Fat digestion and absorption	0.037
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.037
592	alpha-Linolenic acid metabolism	0.037
4110	Cell cycle	0.038
COMBINED		
5322	Systemic lupus erythematosus	4.50E-05
4730	Long-term depression	0.002
52	Galactose metabolism	0.0057
4960	Aldosterone-regulated sodium reabsorption	0.0079
5410	Hypertrophic cardiomyopathy (HCM)	0.01
590	Arachidonic acid metabolism	0.012
3030	DNA replication	0.014
5414	Dilated cardiomyopathy	0.015
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.015
4664	Fc epsilon RI signaling pathway	0.017
524	Butirosin and neomycin biosynthesis	0.019
4320	Dorso-ventral axis formation	0.022

604	Glycosphingolipid biosynthesis - ganglio series	0.029
4912	GnRH signaling pathway	0.029
591	Linoleic acid metabolism	0.033
450	Selenocompound metabolism	0.034
4662	B cell receptor signaling pathway	0.037
130	Ubiquinone and other terpenoid-quinone biosynthesis	0.037
592	alpha-Linolenic acid metabolism	0.04

Table SI 6.5b: KEGG pathway enrichment categories for the genes in the top 5% of Neandertal ancestry corrected for B-statistic. The raw p-values are given for Europeans, East Asians and a combination of Europeans and East Asians. Categories identified as significant in all populations are in bold.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Falcon, S. and R. Gentleman (2007). "Using GOstats to test gene lists for GO term association." Bioinformatics **23**(2): 257-258.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Orr, H. A. and M. Turelli (2001). "The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities." Evolution **55**(6): 1085-1094.

Prufer, K., B. Muetzel, et al. (2007). "FUNC: a package for detecting significant associations between gene sets and ontological annotations." BMC bioinformatics **8**: 41.

Prüfer, K., F. Racimo, et al. (2013). "The complete genome sequence of a Neandertal from the Altai Mountains." submitted.

Pruitt, K. D., J. Harrow, et al. (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes." Genome research **19**(7): 1316-1323.

Ross, M. T., D. V. Grafham, et al. (2005). "The DNA sequence of the human X chromosome." Nature **434**(7031): 325-337.

Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proceedings of the National Academy of Sciences of the United States of America **101**(16): 6062-6067.

SI 7 Neandertal derived alleles segregating in modern humans and their association to phenotype

In this section, we attempt to identify alleles segregating in modern humans that were introduced through Neandertal gene flow and to associate these alleles with phenotypes that have been studied in GWAS. *A priori*, we expect GWAS to have low power to detect association with variants that are Neandertal derived since the allele frequencies of these variants tend to be quite low $< 10\%$. Further, the causal variant may often not be included on genotyping arrays. Despite the low power, we expect that this analysis might give us a first look at the phenotypic impact of Neandertal gene flow.

Firstly, we identified whether an allele segregating in a target population likely owes its origin to the Neandertal gene flow.

- We identified sets of alleles that are confidently labeled as Neandertal, $\mathcal{N}$, by scanning for alleles with marginal probability of Neandertal ancestry ≥ 0.90. We also identified sets of alleles that are confidently labeled as non-Neandertal, $\mathcal{MH}$, by scanning for alleles with marginal probability ≤ 0.1).
- For each SNP called in the 1000 Genomes Project dataset that pass our filters, we required that none of the derived alleles at this SNP falls on one of the alleles in the set $\mathcal{MH}$ and all of the alleles in $\mathcal{N}$ carry the derived allele. This procedure allows for some false negatives in the predictions of the CRF.
- We also required that the derived allele is absent in the panel of 176 YRI.

We ran this procedure on the combined calls from the European and East Asian populations. This procedure yielded a total of 97365 SNPs that are likely to be *Neandertal-derived*.

To associate these Neandertal-derived alleles to phenotype, we downloaded the variants listed in the NHGRI GWAS catalog (Hindorff et al., 2009) (we used a downloaded version that had its latest entry dated to 04/05/2013). We only retained entries for which the reported association is a SNP, the SNP has been assigned a rsid and for which the nominal P-value is $< 5 \times 10^{-8}$. This procedure resulted in 5022 entries of genomewide significant associations.

We then intersected the Neandertal-derived SNPs with the SNPs in the filtered NHGRI GWAS catalog (we did not expand this set to include other SNPs that might be in high LD to the SNPs in our set). 6 of the Neandertal-derived SNPs were present in the GWAS catalog (Extended Data Table 2).

In addition to the GWAS catalog, we reexamined four non-synonymous SNPs that were found to tag the risk haplotype for type 2 diabetes in a recent GWAS in Latinos (The SIGMA Type 2 Diabetes Consortium). Analyses of the geographic distribution, divergence and the genetic length of this haplotype as well as comparison to the high-coverage Neandertal genome (Prüfer et al., 2013) suggested that this haplotype is introgressed from Neandertals. We were interested to see if any of the four SNPs on the risk haplotype were predicted to be Neandertal-derived according to our criteria. Three of the four SNPs (at positions 6945087,6945483 and 6946330 on chromosome 17) were found to be Neandertal-derived according to our criteria. The fourth (6946287 on chromosome 17) is present at appreciable frequencies in sub-Saharan Africans and is not predicted to be Neandertal-derived.

Some notes on the Neandertal derived alleles listed in Extended Data Table 2 are as follows.

1. rs12531711 has been shown to be associated with primary biliary cirrhosis (Mells et al, 2011). It has also been shown to be associated with systemic lupus erythematosus (SLE) both in cases for which the anti-dsDNA autobodies were observed and those that were not (Chung et al.,

2011). The catalog also lists a meta-analysis of GWAS in European populations that found evidence of association between this SNP and SLE (Table 2, (Lee et al., 2012)). According to the catalog, the predictor allele in this study is A. However, the reported odds ratio would then be inconsistent with the odds ratio computed using Chung et al. (2011) which uses G as the predictor allele. We think that the latter is correct based on the allele frequencies and reported odds ratios *i.e.*, allele G increases the risk for SLE at this SNP.

2. rs3025343 : Derived allele is negatively associated with smoking cessation (Furberg et al., 2010).

3. rs7076156 :Derived allele is risk-decreasing in Ashkenazi Jews (Kenny et al., 2012).

4. rs12571093 : Derived allele is associated with decreased disk area (Macgregor et al., 2010).

5. rs1834481 : Derived allele is associated with decreased plasma IL-18 (He et al., 2010).

6. rs11175593 : Derived allele is risk-increasing (Barrett et al., 2008).

SI 8 Analysis of genomic regions deficient in Neandertal ancestry

Extended Data Fig.2 suggests that there exist several large regions of the genome that have little Neandertal ancestry.

SI 8.1 Identification of regions deficient in Neandertal ancestry

To assess the existence of regions deficient in Neandertal ancestry in a robust manner, we measured the fraction of Neandertal ancestry $ta_t(w)$ that exceeds a threshold t averaged across all SNPs and individuals within non-overlapping windows that tile the genome:

$$ta_t(w) = \frac{\sum_{s=1}^{m}\sum_{j=1}^{n}\mathbf{1}\{\gamma_{s,j} > t\}}{m|\{j \in w\}|} \tag{8}$$

Here $t \in [0,1]$ is a threshold. A SNP in an individual is declared to be Neandertal derived if the marginal probability assigned by the CRF to the SNP exceeds this threshold. For most analyses, we have chosen $t = 0.9$. While this choice resulted in Neandertal ancestry calls that have low false discovery rates, the recall is reduced at this threshold. To assess whether a window is deficient in Neandertal ancestry however, a high recall is desirable. Hence, we choose $t = 0.25$. Our analysis of the empirical precision and recall show that, at this threshold, both the precision and recall are high ($> 80\%$). Further, we assess deficiency of Neandertal ancestry in large windows ($w = 10$ Mb) – this averages the effect of drift later in gene flow and makes the observations less likely to be an artifact of reduced power.

We excluded all windows that overlap (over any part of their length) the centromeres or the telomeres. We further restricted our analysis to windows in which the number of SNPs that pass filters is at least 1000. We also discarded windows for which the genetic length of the window was less than 1.96 standard deviations from the mean – a reduced Neandertal ancestry in these windows is expected to be more variable due to the smaller number of recombined Neandertal haplotypes.

In this set of 227 filtered windows, the average number of SNPs is $99,036$ with a standard deviation of $14,030$ and a range of $12,190$ to $150,453$. We measured the distribution of $ta_{0.25}$ in this set. The mean and standard deviation of $ta_{0.25}$ measured in EUR are 1.79% and 1.45%. We observe that four windows have a $ta_{0.25}$ less than 0.1% with one window having $ta_{0.25} = 0$. In ASN, the mean and standard deviation of $ta_{0.25}$ are 2.25% and 2.27% with fourteen windows $< 0.1\%$ with one window having $ta_{0.25} = 0$. Amongst windows with $ta_{0.25} < 0.1\%$, there is one window, $7 : 110,000 - 000 - 120,000,000$ that has $ta_{0.25} = 0$ and is common to EUR and ASN. We also examined chromosome X for 10 Mb windows with $ta_{0.25} < 0.1\%$. We detected five and three windows in Europeans and East Asians respectively.

To assess if regions deficient in Neandertal ancestry might be artifacts caused by reduced power to detect Neandertal ancestry, we computed Spearman's correlation coefficient ρ between Neandertal ancestry as measured by $ta_{0.25}$ and the genetic length of 10 Mb windows. If the regions deficient in Neandertal ancestry are artifacts of reduced power, we would expect the proportion of Neandertal ancestry in them to be negatively correlated with recombination rate as Neandertal haplotypes would be longer and thus easier to detect in regions of low recombination rate, but we instead see a significant trend in the positive direction ($\rho = 0.221$ in EUR; $P = 4.4 \times 10^{-3}$ and $\rho = 0.226$ in ASN; $P = 1.92 \times 10^{-3}$).

There are two possible explanations for these large regions with relatively little Neandertal ancestry:

	Population		la		$ta_{0.25}$		$ta_{0.9}$
		ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$
Autosomes	EUR	0.32 (0.0162)	86.310	0.131 (0.0169)	14.064	0.0394 (0.0162)	1.822
Autosomes	ASN	0.305 (0.0175)	67.411	0.0963 (0.0202)	5.698	0.0209 (0.0198)	0.533
X	EUR	0.276 (0.0809)	3.358	0.21 (0.0639)	3.150	0.15 (0.0466)	3.124
X	ASN	0.176 (0.105)	1.062	0.111 (0.105)	0.574	0.107 (0.0972)	0.617

Table SI 8.1: Relationship between Neandertal ancestry and B-statistic. ρ refers to Spearman's correlation coefficient.

- Regions that have drifted to low frequencies early on since gene flow.
- Rapid selection against the introgressing Neandertal haplotype.

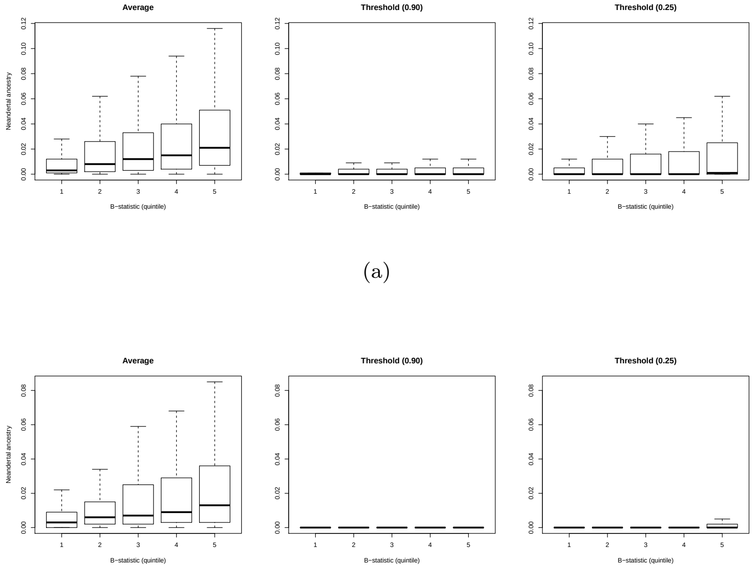
SI 8.2 Correlation of Neandertal ancestry with B-statistics is consistent with the action of selection on Neandertal alleles

We analyzed if the proportion of Neandertal ancestry in a genomic region is correlated to the B-statistic, a measure of the strength of background selection (McVicker et al., 2009). B-statistic is reduced in the vicinity of genes and other functional elements. We analyzed the relationship between Neandertal ancestry and the B-statistic at different size scales, on the X chromosome and the autosomes and in Europeans and East Asians.

B-statistics were lifted over to hg19 coordinates. We then annotated each of the SNPs that we analyzed with the B-statistic of the genomic region in which the SNP falls. In our first analysis, we partitioned SNPs into quintiles based on their B-statistic annotation. At each SNP, we considered several estimates of the Neandertal ancestry : la which computes the average over the marginal probability of Neandertal ancestry assigned to each individual haplotype, $ta_{0.9}$ which computes the average fraction of alleles across individuals that attain a marginal probability of ≥ 0.90 and $ta_{0.25}(w)$ that computes the analogous statistic for a threshold of 0.25. Figures SI 8.1 and SI 8.2 plot the relationship between Neandertal ancestry and B-statistic quintile. We observe a trend of the median Neandertal ancestry increasing with the quintile *i.e.*, the median Neandertal ancestry is higher in regions of high B. This relation is strongest for the la statistic relative to either of the thresholded statistics. A likely reason for this difference is that power to detect Neandertal ancestry is lower in regions of high B (Section SI 2.4). Using a constant threshold across the genome is expected to reduce this signal.

We estimated Spearman's correlation coefficient ρ between Neandertal ancestry and B-statistic (Table SI 8.1). We performed a block jackknife in 10 Mb windows to estimate the standard error of ρ. We see a statistically significant correlation between B-statistic and different summaries of the Neandertal ancestry (except in the case of East Asians and the $ta_{0.90}$ statistic which is likely to be due to its decreased power).

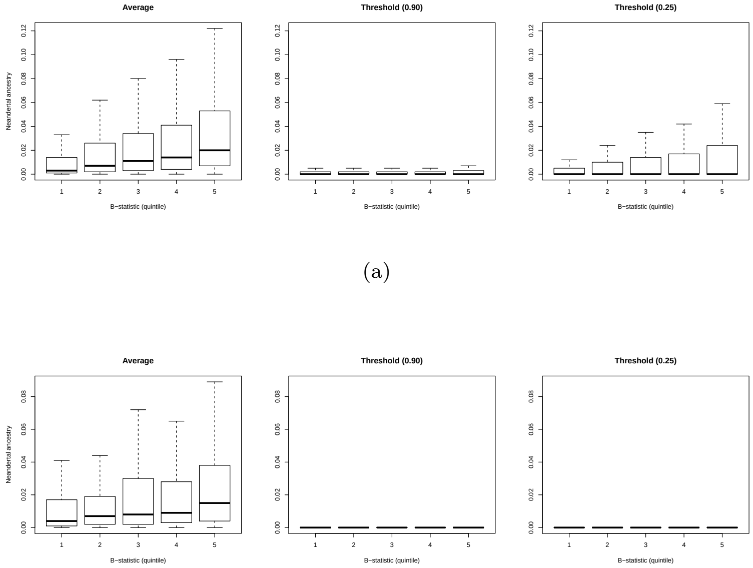
We decided to interrogate the relation between Neandertal ancestry and B at several size scales as such an analysis can potentially provide insights into the strength of selection. We partitioned the genome into non-overlapping windows of size w. Within each window w, we estimated the



(a)

(b)

Figure SI 8.1: Neandertal ancestry vs quintile of B-statistic in EUR on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$.



(a)

(b)

Figure SI 8.2: Neandertal ancestry vs quintile of B-statistic in ASN on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$.

Neandertal ancestry. We considered several estimates of the Neandertal ancestry : $la(w)$ which computes the average over the marginal probability of Neandertal ancestry assigned to each SNP on each individual haplotype within window w, $ta_{0.9}(w)$ which computes the average fraction of alleles across individuals and SNPs within window w that attain a marginal probability of ≥ 0.90 and $ta_{0.25}(w)$ that computes the analogous statistic for a threshold of 0.25. We further restricted our analysis to windows in which the number of SNPs that pass filters is at least 10 (at such windows, our estimate of the proportion of Neandertal ancestry is likely to be noisy). We did not explicitly remove windows that overlap the centromeres or telomeres as this would impact the analysis at at large size scale disproportionately. Analogous to the way we estimated Neandertal ancestry, we computed the average value of the B-statistic over each window w by averaging the B-statistic assigned to each SNP within the window.

Figures SI 8.3, SI 8.5, SI 8.7 and SI 8.10 plot the relationship between Neandertal ancestry and B-statistics for varying window sizes. We estimated Spearman's correlation coefficient ρ between Neandertal ancestry and B-statistic (Table SI 8.2 and SI 8.3). We performed a block bootstrap to resample the windows. We used 1000 bootstrap replicates to estimate the standard error of ρ and to construct a Z-score which we then use to assign a P-value to the two-sided hypothesis that ρ is zero.

We make several observations from Tables SI 8.2 and SI 8.3:

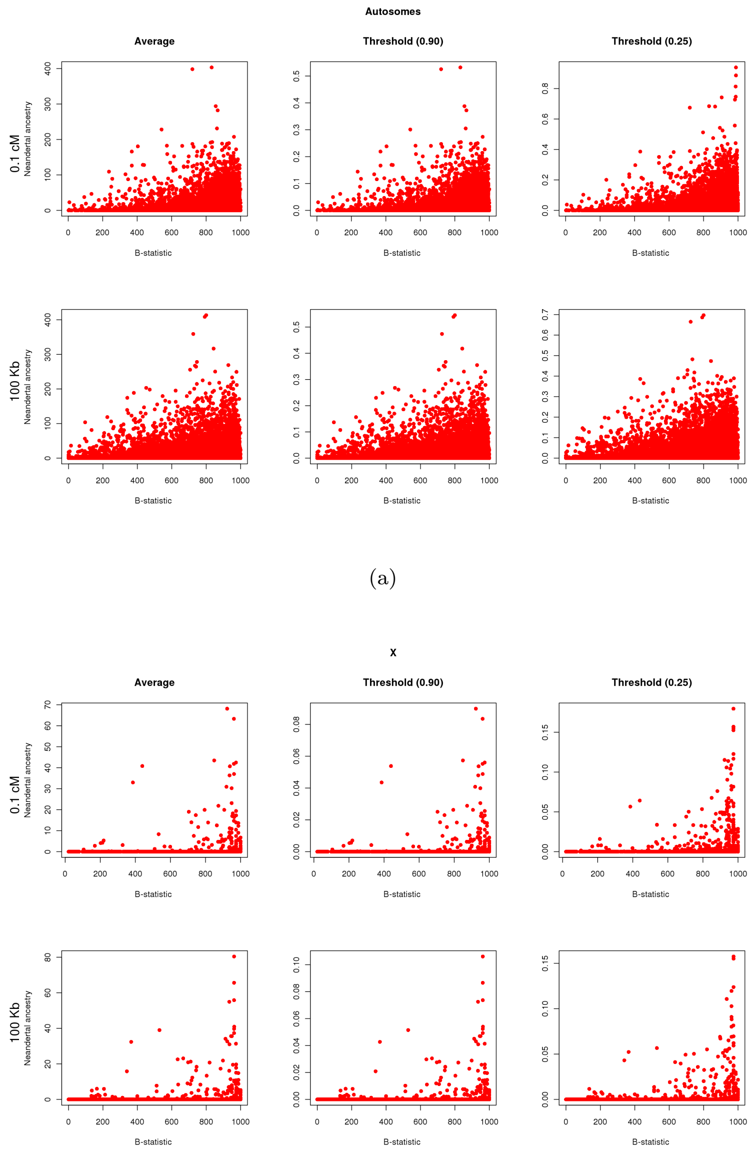
- At a 100 Kb size scale, Neandertal ancestry is positively correlated with B-statistic. This relationship holds on the autosomes and the X chromosome, in Europeans and East Asians, and holds for all the summaries of Neandertal ancestry considered (the correlation being stronger on the autosomes than the X and in Europeans than in East Asians). The relationship is statistically significant in all cases.

- At a 10Mb size scale, we see the same qualitative relationship as at a 100 Kb size scale although the strengths of the correlation are reduced due to a smaller number of windows at this scale. The relationship is stronger in Europeans than in East Asians and is not statistically significant in East Asians though consistent with the signal at a 100 Kb scale.

- At a 0.1 cM size scale, we see a similar trend on the autosomes. An exception is that Neandertal ancestry as measured by $ta_{0.90}$ is negatively correlated with B-statistic on the autosomes. This effect reverses for the $ta_{0.25}$ and the la statistics. One possible contributor to this effect is that the power to identify Neandertal haplotypes is reduced in regions of high B-statistic, as genetic diversity is higher. Thus the correlation of $ta_{0.90}$ is expected to be an underestimate of the true slope.

SI 8.3 Variation of Neandertal ancestry at 10 Mb size scales is consistent with drift early after introgression

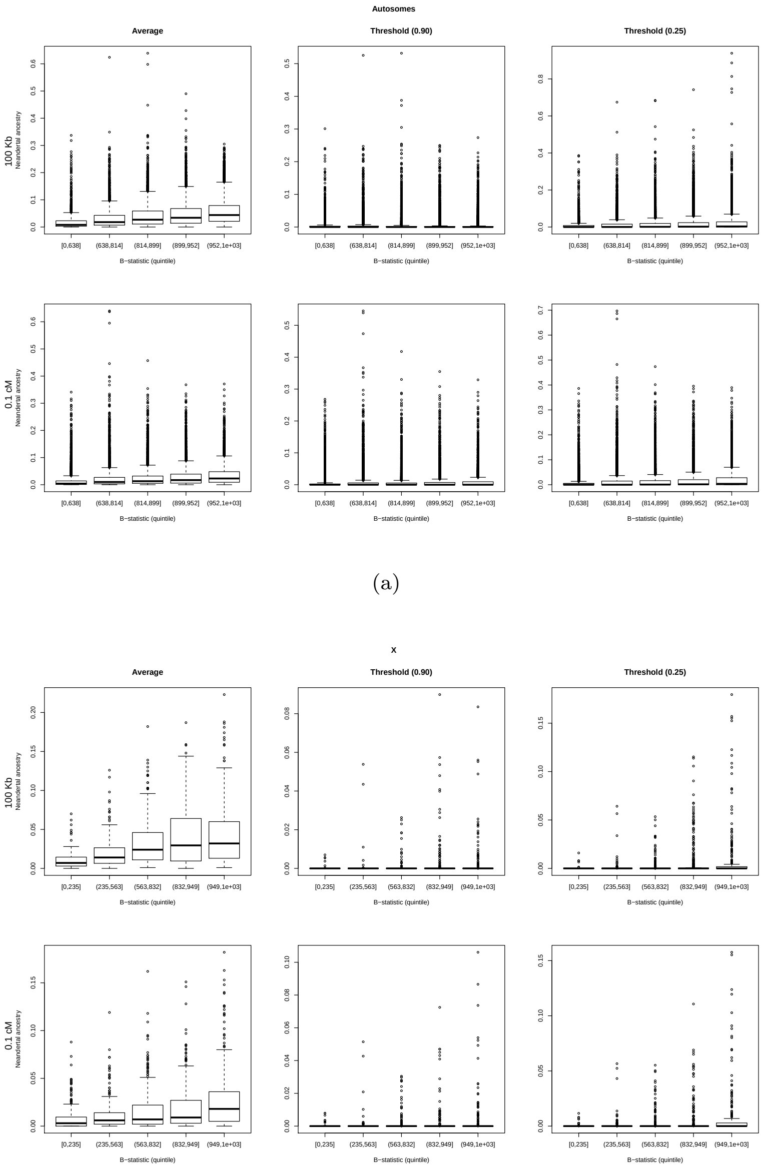
To understand the contribution of demographic effects to the variation in Neandertal ancestry, we use the idea that if drift is responsible for the observed pattern, the entire distribution of Neandertal ancestry should be affected. Selection that is localized to a small number of regions of the genome, on the other hand, while producing an excess of low-frequency regions, should have little effect on the bulk of the distribution. To test this, we measured the coefficient of variation of $ta_{0.25}$ at a 10 Mb scale. We measure a cv of 0.808 ± 0.044 in EUR and 1.010 ± 0.056 in ASN (standard errors were obtained using a block jackknife where a block consists of a 10 Mb window). We also measured the cv in constant 10 cM windows. We observed a cv of 0.838 ± 0.034 in EUR and 1.060 ± 0.048 in ASN, consistent with the findings from physical distance-based windows. Figure SI 8.11 shows the cv of



(a)

(b)

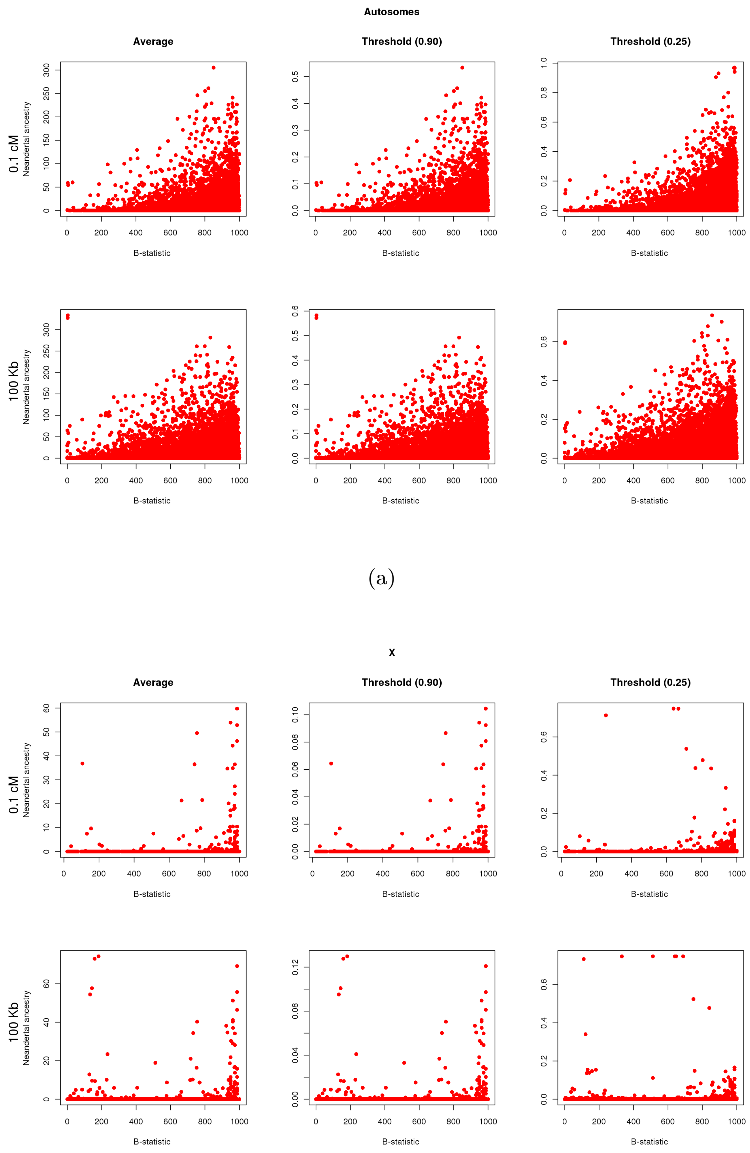
Figure SI 8.3: Neandertal ancestry vs B-statistic in EUR on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (0.1 cM) while the bottom row plots windows that are of constant physical size (100 Kb).



(a)

(b)

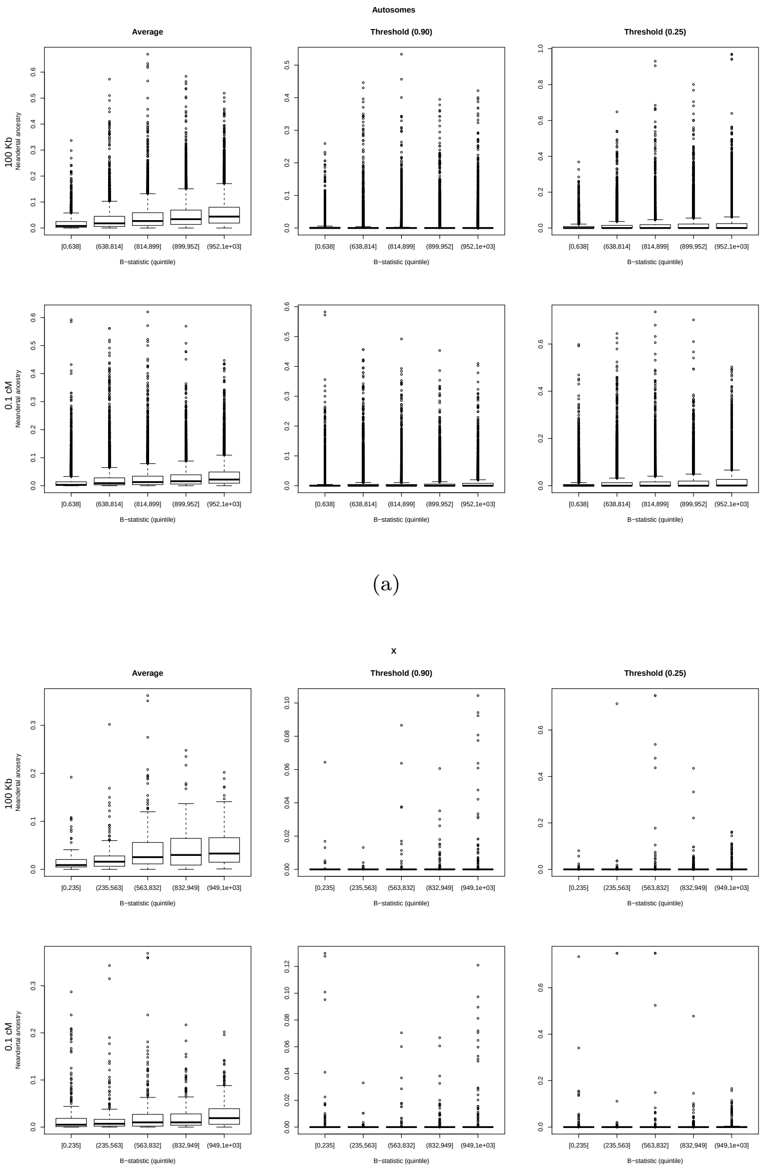
Figure SI 8.4: Neandertal ancestry vs quintile of B-statistic in EUR on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (0.1 cM) while the bottom row plots windows that are of constant physical size (100 Kb).



(a)

(b)

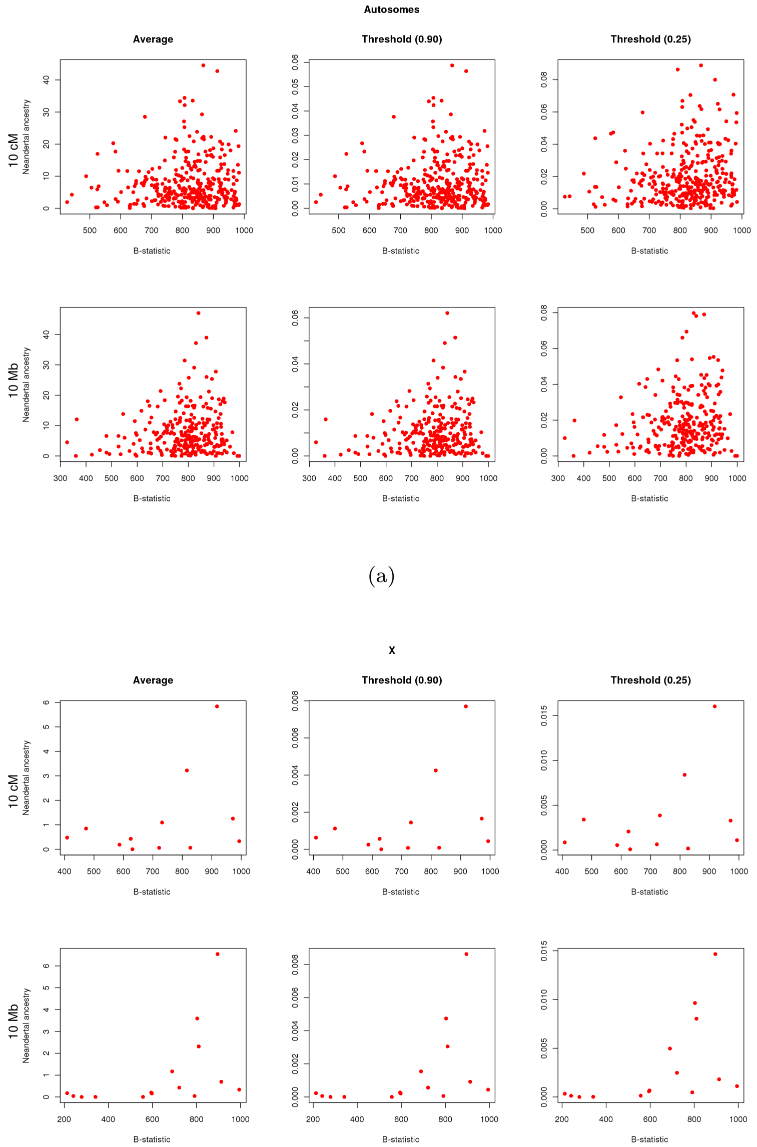
Figure SI 8.5: Neandertal ancestry vs B-statistic in ASN on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (0.1 cM) while the bottom row plots windows that are of constant physical size (100 Kb).



(a)

(b)

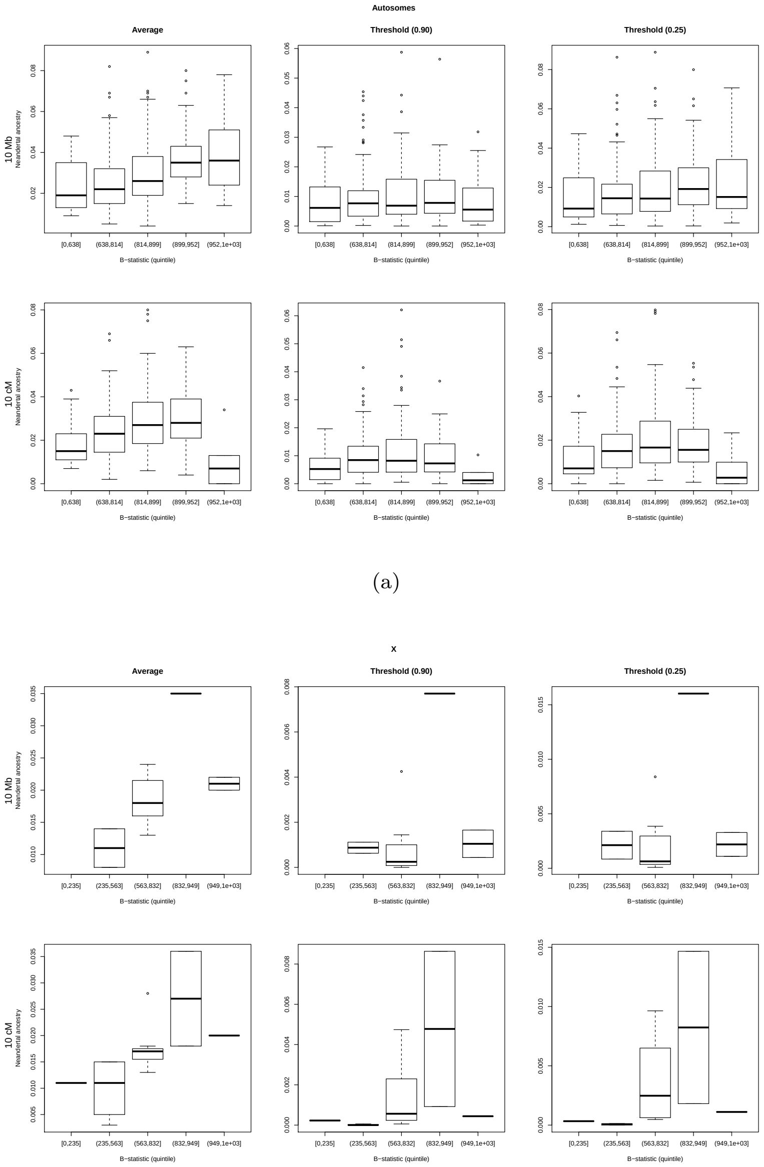
Figure SI 8.6: Neandertal ancestry vs quintile of B-statistic in ASN on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (0.1 cM) while the bottom row plots windows that are of constant physical size (100 Kb).



(a)

(b)

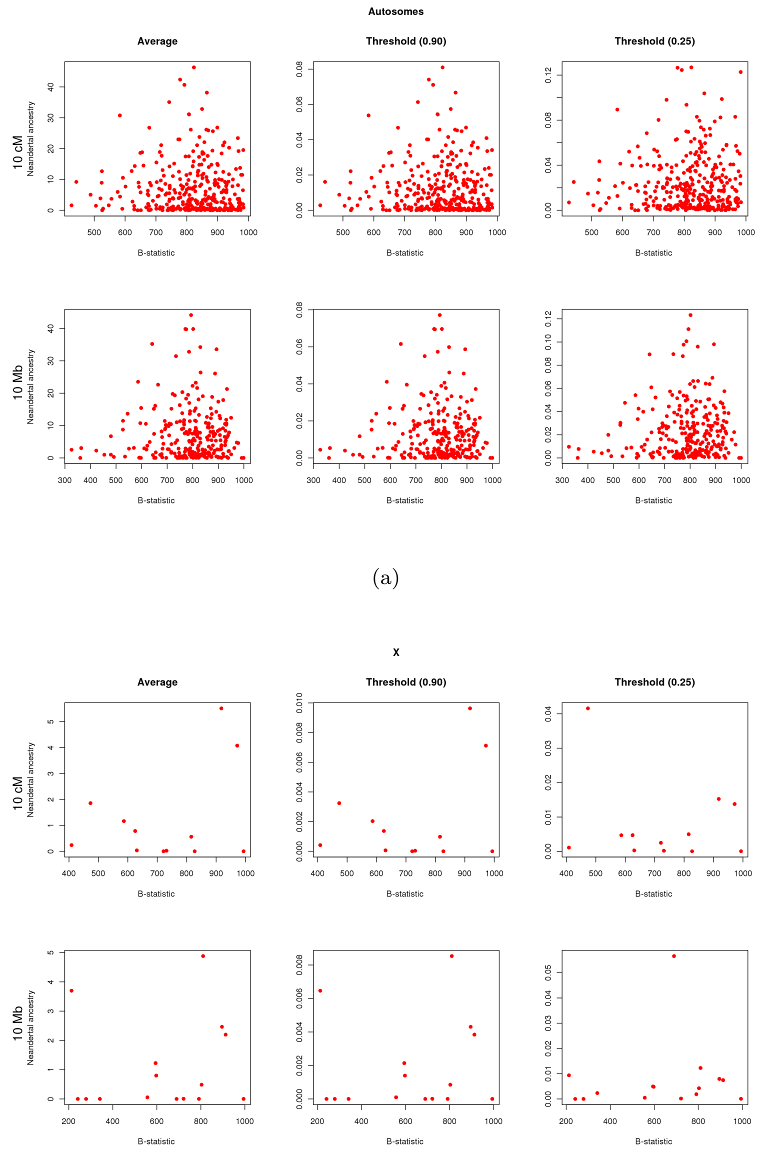
Figure SI 8.7: Neandertal ancestry vs B-statistic in EUR on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (10 cM) while the bottom row plots windows that are of constant physical size (10 Mb).



(a)

(b)

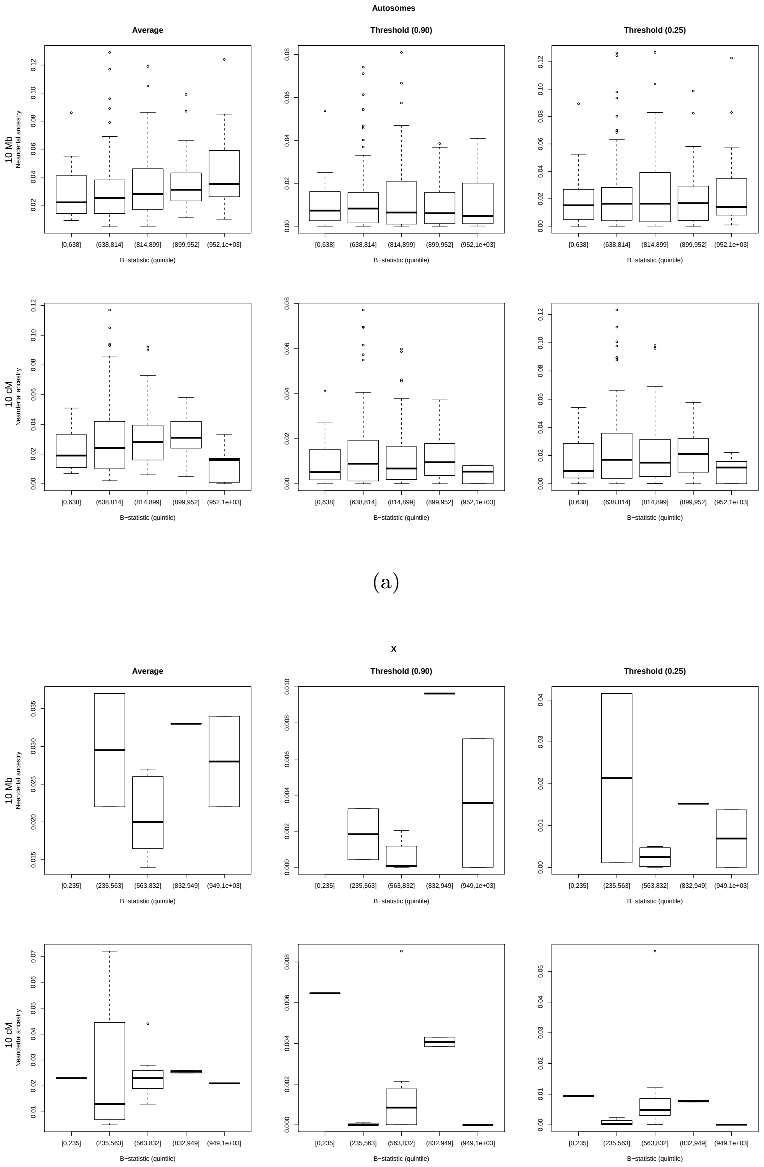
Figure SI 8.8: Neandertal ancestry vs quintile of B-statistic in EUR on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (10 cM) while the bottom row plots windows that are of constant physical size (10 Mb).



(a)

(b)

Figure SI 8.9: Neandertal ancestry vs B-statistic in ASN on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (10 cM) while the bottom row plots windows that are of constant physical size (10 Mb).



(a)

(b)

Figure SI8.10: Neandertal ancestry vs quintile of B-statistic in ASN on the a) autosomes and the b) X-chromosome . In each panel a) and b), each column plots a different summary of the Neandertal ancestry within a window. The leftmost plots la, the middle plots $ta_{0.9}$ and the rightmost $ta_{0.25}$. The top row in each panel plots this relationship for windows that are of constant genetic length (10 cM) while the bottom row plots windows that are of constant physical size (10 Mb).

	Population	la		$ta_{0.25}$		$ta_{0.9}$	
		ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$
Autosomes	EUR	0.327 (0.00566)	727.812	0.167 (0.006)	170.332	0.0979 (0.00633)	53.218
Autosomes	ASN	0.313 (0.00582)	630.638	0.129 (0.0059)	105.429	0.0703 (0.00617)	29.354
X	EUR	0.314 (0.0243)	37.387	0.238 (0.0228)	24.680	0.201 (0.0224)	18.569
X	ASN	0.2 (0.026)	13.740	0.138 (0.0258)	7.064	0.108 (0.0269)	4.239
Autosomes	EUR	0.303 (0.00601)	553.922	0.111 (0.00637)	67.123	-0.029 (0.00638)	5.243
Autosomes	ASN	0.279 (0.00621)	439.239	0.0575 (0.00647)	18.162	-0.0523 (0.00646)	15.271
X	EUR	0.219 (0.0277)	14.564	0.148 (0.0263)	7.711	0.0827 (0.0272)	2.621
X	ASN	0.177 (0.028)	9.548	0.0946 (0.0268)	3.387	0.0632 (0.0285)	1.575

Table SI 8.2: Relationship between Neandertal ancestry and B-statistic at a 100 Kb and at a 0.1 cM size scale. ρ refers to Spearman's correlation coefficient.

	Population	la		$ta_{0.25}$		$ta_{0.9}$	
		ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$	ρ (se)	$-log_{10}(pval)$
Autosomes	EUR	0.239 (0.0588)	4.328	0.137 (0.0599)	1.651	0.0841 (0.0621)	0.755
Autosomes	ASN	0.147 (0.0569)	2.002	0.0863 (0.0568)	0.891	0.0506 (0.0592)	0.406
X	EUR	0.741 (0.154)	5.801	0.768 (0.128)	8.646	0.685 (0.134)	6.452
X	ASN	0.193 (0.268)	0.327	0.227 (0.291)	0.361	0.207 (0.318)	0.287
Autosomes	EUR	0.333 (0.048)	11.399	0.143 (0.0519)	2.241	0.0411 (0.0534)	0.356
Autosomes	ASN	0.193 (0.0504)	3.880	0.00735 (0.053)	0.051	-0.0525 (0.0526)	0.498
X	EUR	0.674 (0.179)	3.771	0.266 (0.264)	0.502	0.259 (0.282)	0.445
X	ASN	-0.0808 (0.312)	0.099	-0.161 (0.344)	0.194	-0.127 (0.356)	0.142

Table SI 8.3: Relationship between Neandertal ancestry and B-statistic at a 10 Mb and at a 10 cM size scale. ρ refers to Spearman's correlation coefficient.

$ta_{0.25}$ in CEU and CHB at varying physical and genetic lengths. Figure SI 8.13 shows the the cv of the $ta_{0.90}$ statistic which is also qualitatively similar to the behavior of the $ta_{0.25}$ statistic.

To see if the cv might be explained by a demographic model, we simulated data under several demographic models. We used a procedure similar to the one described in Section SI 2.3. For each simulated dataset, we ran the CRF with model parameters as estimated in Section SI 2.1. We then estimated the cv of the $ta_{0.25}$ statistic in 10 cM windows.

Since there is considerable uncertainty in the demographic model relating modern humans and Neandertals, we started with a reasonable demographic model and perturbed each parameter of the model in turn. We simulated 100 European haplotypes, 100 African haplotypes as well as a single Neandertal haplotype. We simulated 300 1 Mb regions using a version of ms (Hudson, 2002). However, since we are interested in the Neandertal ancestry at a 10 cM size scale, we set the mutation rate and recombination rates to 10 times their genomewide average *i.e*, $\mu = 1.2 \times 10^{-7}, r = 1.3 \times 10^{-7}$ per bp per generation. Using these parameters, a 10 cM window has a physical length of 769 kb. In each set of simulations, we varied each parameter of the demographic model in turn over a grid of size 6 that includes the endpoints (as described in Section SI 2.3).

1. T_1 – the time of split of Africans and Europeans. Default of 2500 generations. Varied from 2500 to 5000.

2. T_2 – the time of split of modern humans and Neandertals. Default of 12000 generations. Varied from 9000 to 13000.

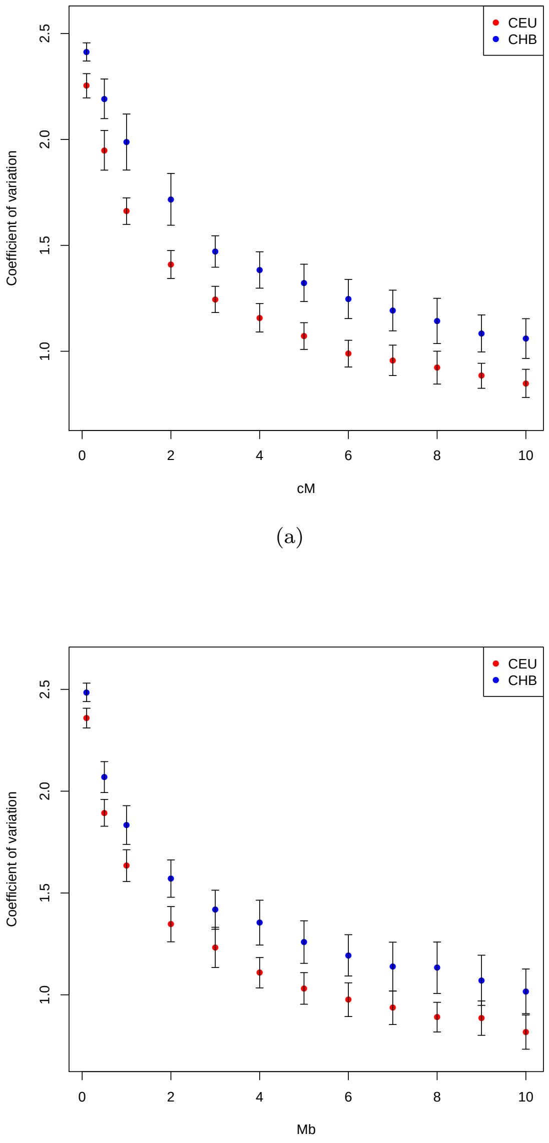
3. T_{GF} – the time of Neandertal gene flow. Default of 2000 generations. Varied from 500 to 3000.

4. N_1 – the effective population size (assumed constant) in Europeans since gene flow. Default of 10000. Varied from 5000 to 50000.

We set the admixture proportion to 0.02 which is consistent with the most recent estimates from formal methods (Prüfer et al., 2013). We also assumed a bottleneck in the Neandertals beginning 6120 generations ago and ending 6000 generations ago, in which the Neandertal effective population size is reduced to 100 as was done in Section SI 1.

In these simulations, the drift since Neandertal gene flow ranges from 0.025 to 0.2. These settings span the estimates of drift since Neandertal gene flow in European and East Asian populations. One common feature of all these models is that Neandertal gene flow occurred into a large population. We ran the CRF on each simulated dataset with parameters estimated on the demographic model described in Table SI 2.1 in Section SI 2.1. We find that the maximum coefficient of variation at a 10 cM size scale across models is 0.54 ± 0.02.

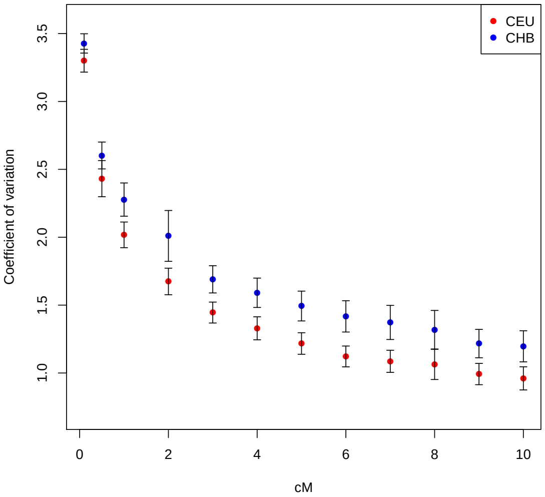
We introduced a bottleneck of strength 0.10 (the bottleneck duration was fixed to 20 generations and the effective population size was fixed to 100) at various times both before and after the gene flow event (ranging from 1020 to 2520 generations in a grid of size 6). For bottlenecks that occurred before the gene flow, the coefficient of variation is 0.52 ± 0.024. For bottlenecks that occurred after the first 300 generations, the coefficient of variation is 0.75 ± 0.024. However, a bottleneck that starts 80 generations after the time of gene flow produces a coefficient of variation of 1.03 ± 0.05 while a bottleneck that start 180 generations after gene flow has a cv of 0.86. These results are consistent with a model in which there is increased drift early on in the history of non-African populations since Neandertal gene flow. One such model requires Neandertal gene flow into a relatively small population (in terms of the effective population size). However, it is also plausible that the Neandertal gene flow was into a large population which had a reduced effective population size soon after. This analysis also does not distinguish if all or some of this drift could have been shared between the European and East Asian populations. For example, we find only 1 window



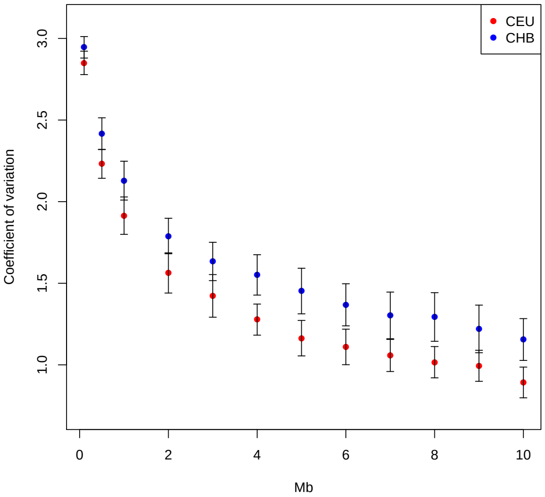
(a)

(b)

Figure SI 8.11: Coefficient of variation of Neandertal ancestry as estimated by $ta_{0.25}$ in CEU and CHB in windows of a) constant genetic size and b) constant physical size.



(a)



(b)

Figure SI 8.12: Coefficient of variation of Neandertal ancestry as estimated by $ta_{0.90}$ in CEU and CHB in windows of a) constant genetic size and b) constant physical size.

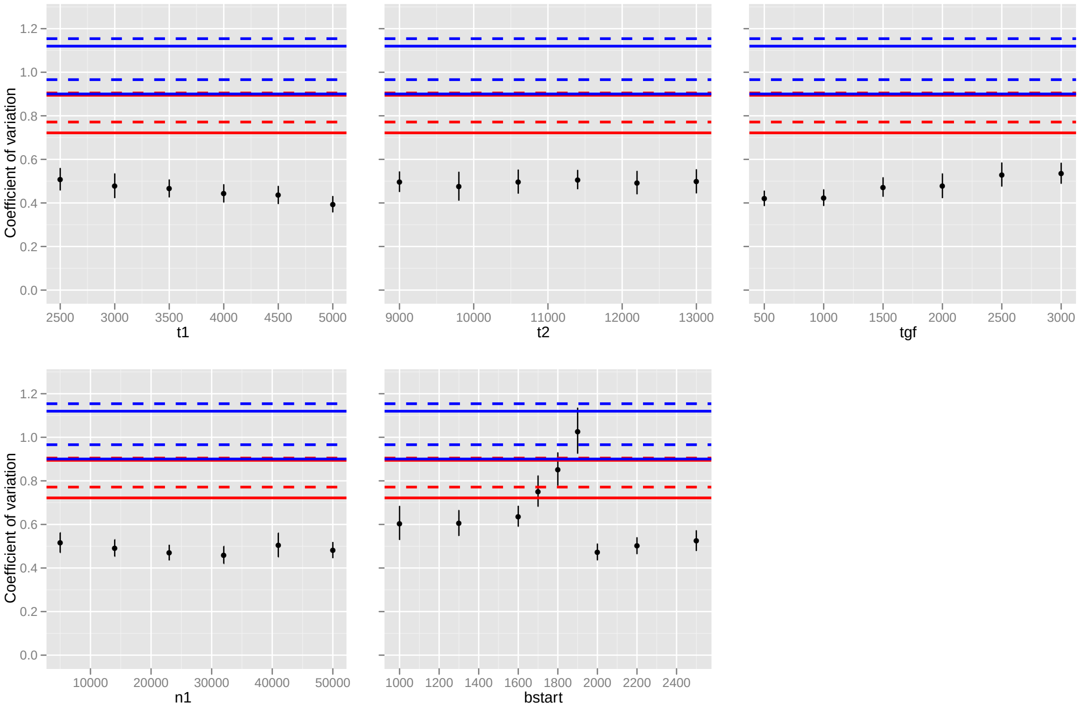


Figure SI 8.13: Coefficient of variation of Neandertal ancestry as estimated by $ta_{0.25}$ on simulated data. The proportion of Neandertal gene flow was set to 0.02. The x-axis denotes a single parameter of the demographic model that was varied while keeping all other parameters fixed (see Section SI 8.3). We plot the point estimate and 1.96× the standard errors. Solid red (blue) lines are the 95% CIs of the coefficient of variation in EUR (ASN) at a 10 Mb size scale. Dashed red (blue) lines denote the 95% CIs at a 10 cM size scale in EUR (ASN).

overlapping between Europeans and East Asians at a frequency $< 0.1\%$ (a caveat however is that it is not clear that we ought to be using a common threshold for both populations given their differing means and variances).

In introducing a bottleneck, our analysis has been conservative in assuming that the total drift on the non-African population is at least 0.10. If we further constrain this quantity based on estimates of say 0.07 in Europeans, this might constrain the drift to occur even earlier in the history of gene flow.

SI 9 Unbiased statistics support enhanced Neandertal ancestry in gene-poor regions

SI 9.1 Motivation

In Figure 3 of the main text, we show that the proportion of the genome that is inferred by our method to be of Neandertal ancestry is lower in gene-dense regions in both Europeans and East Asians.

In the main text, we interpret this signal as evidence that Neandertal ancestry has been removed from gene-dense regions (as measured by a low value of the B-statistic) via the action of natural selection. However, an alternative interpretation of these patterns is that these observations do not reflect true biology, but instead are artifacts of the fact that our local ancestry inference method has varying sensitivity to detecting true segments of Neandertal ancestry depending on underlying genomic features.

As a first way of exploring this alternative possibility, in Section SI 2.4, we report simulations that show that in regions of the genome with a low time since the most recent common ancestor (typical for regions with low B-statistics), we expect to have increased power to detect Neandertal ancestry. This makes our observation of a reduced rate of Neandertal ancestry in these regions all the more surprising as if our signal was a result of varying power, the local Neandertal ancestry in low B-statistic regions ought to be increased, not reduced as we observe.

As a second way of exploring the possibility that our signal might be an artifact of variation in the power of our method across the genome, in this note we estimate Neandertal ancestry as a function of B-statistic using an ancestry estimation statistic that is not biased by features that vary spatially across the genome such as mutation rate, recombination rate, and average time since the most recent common ancestor. While this statistic gives noisier estimates than local ancestry inference, the fact that it is unbiased allows us to carry out formal tests.

SI 9.2 Method

We analyzed data from 27 deeply sequenced genomes: 25 genomes from present-day humans and the high coverage Altai Neandertal and Denisova genomes. In each deeply sequenced sample, we restricted to nucleotides that passed the stronger of the two sets of filters described in SI 5 of Prüfer et al. (2013) (Map35_100%), and further required a genotype quality scores of GQ≥45 at each analyzed site. We finally required that we could determine the ancestral allele based on comparison to chimpanzee and at least one of two great ape genomes (gorilla or orangutan).

To increase power for our analysis, we analyzed the following pools of present-day humans, restricting to nucleotides that had a genotype in at least one sample in the specified pool.

"European"	n=4	2 French, 2 Sardinian
"Eastern"	n=7	2 Han, 2 Dai, 2 Karitiana, 1 Mixe
"Non-African"	n=11	2 French, 2 Sardinian, 2 Han, 2 Dai, 2 Karitiana, 1 Mixe
"African"	n=6	2 Dinka, 2 Yoruba, 2 Mbuti

We divided the genome into quintiles of B-statistic, each containing an approximately equal amount of data. In each quintile, we computed the following unbiased estimator of Neandertal ancestry first presented in SI 8 of Reich et al. (2010) :

$$\hat{R} = \frac{\hat{S}(Non-African, African, Denisova, Chimpanzee)}{\hat{S}(Neandertal, African, Denisova, Chimpanzee)}$$

This statistic measures the excess rate of matching of a non-African sample to Denisova using Africans as a baseline, compared with what is seen for a 100% Neandertal (Altai). This estimate is unbiased regardless of the mutation rate, recombination rate, and average time since the most recent common ancestor, as these are expected to affect the numerator and denominator equally.

We computed Neandertal ancestry estimate for each quintile as a fraction of the average:

$$\hat{Y}_{quintile} = \frac{\hat{R}_{quintile}}{\hat{R}_{whole\ genome}}$$

We obtained a standard error using a Block Jackknife by dividing the genome into 100 equally sized contiguous chunks and studying the variation across chunks (Kunsch, 1989).

SI 9.3 Results

Pooling data from Europeans and Eastern non-Africans and using both transition and transversion polymorphisms to reduce the standard errors, we find that the proportion of Neandertal ancestry is 1.537 ± 0.152 times larger in the quintile of the genome with the highest B-statistic (B=0.94-1) than in the bottom four quintiles (B=0-0.94) (Extended Data Table 4).

The observed excess is statistically significant at Z=3.82 (nominally $P=6.6\times10^{-5}$ by a one-sided test). To be conservative, we applied a penalty for multiple hypothesis testing (since we visually inspected the data and chose the boundary between bins that maximized the differentiation). We specifically assumed that we tested 10 hypotheses (notionally testing 5 different hypotheses and performing a 2-sided test in each). The value we report in Extended Data Table 4 is thus a P-value with Bonferroni correction for 10 hypotheses tested of P=0.00066.

We also carried out the same analyses in subsets of the data: transversions only (P=0.013), Europeans and transversions only (P=0.0090), and Eastern non-Africans and transversions only (P=0.049). We obtain qualitatively consistent results.

SI 10 Wavelet decomposition of the correlation of Neandertal ancestry across populations

We use wavelet analysis (Percival and Walden, 2005) to assess the correlation between Neandertal ancestry in Europeans and East Asians at different size scales. Intuitively, Neandertal ancestry along the genome can be written as a linear combination of basis functions (constructed by translating and dilating a scaling and wavelet function) that capture the variation of Neandertal ancestry along the genome at different size scales.

To perform wavelet analysis, we consider the $ta_{0.25}$ statistic measured in 10 Kb windows (using 1 Kb windows or using the $ta_{0.90}$ statistic gave similar results). We considered each of the 22 chromosomes individually. The wavelet transform that we apply require the length of the series to be a power of two; hence, we pad the series for each chromosome on the left and the right so that the length is equal to its nearest power of two. Further, there are some windows which have no SNPs and hence a $ta_{0.25}$ statistic of zero. We use a linear interpolation scheme to predict the $ta_{0.25}$ statistic at these windows.

We applied a discrete wavelet transform using Daubechies least-assymetric wavelets with ten vanishing moments (Daubechies, 1988) using the R wavethresh package (wav). This transform decomposes the variation in Neandertal ancestry into size scales of $2^k \times 10$ Kb, $k = 0, \ldots, 11$, the maximum of which corresponds to a size scale that is smaller than the smallest of chromosomes. At each scale k, we computed the Spearman correlation of the detail coefficients for Europeans and East Asians across all chromosomes. To estimate standard errors on the correlation, we used a weighted block jackknife with each chromosome treated as a block and weights set to the number of windows on the chromosome (Busing et al., 1999).

Figure SI 10.1 shows how the correlation in Neandertal ancestry between EUR and ASN at size scales ranging from 10 to $2^{11} \times 10$ Kb. We see significant correlation across all size scales with increasing correlation at a $10 - 20$ Mb scale.

We considered the fact that some of the correlation in Neandertal ancestry proportion that we observe between EUR and ASN is likely to be due to the fact that the sensitivity of our method to detecting Neandertal ancestry is correlated to local genome sequence features that are shared which are in turn shared between EUR and ASN. While this is a real effect, it cannot explain the fact that the correlation grows stronger at larger distance scales, as this effect would in fact predict a decrease in correlation at larger distance scales. Specifically, at larger distance scales we would expect power to detect Neandertal ancestry would not vary much across the genome since we are averaging over diverse sequence features so any correlation that is due to the sensitivity of the method would be expected to decrease.

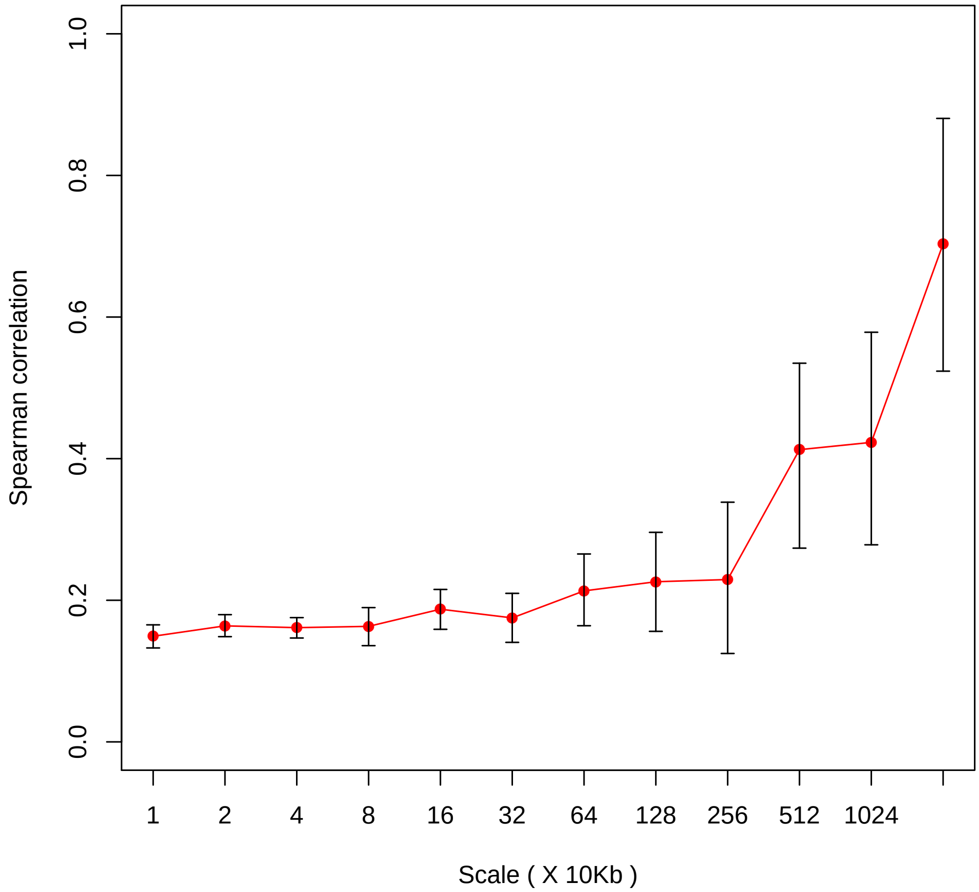


Figure SI 10.1: Correlation of Neandertal ancestry across EUR and ASN at different size scales. Neandertal ancestry was estimated by $ta_{0.25}$ in windows of $10Kb$.

SI 11 Evidence for Neandertal introgression at loci previously identified to be introgressed

To evaluate the power of the CRF on empirical data, we assessed the predicted Neandertal ancestry at loci that have been previously identified to be introgressed (Abi-Rached et al., 2011; Mendez et al., 2012, 2013; Yotova et al., 2011).

We considered all SNPs called by the 1000 Genomes project within each locus. For each locus, we computed two statistics – the first, ta, that only counts alleles at which the predicted Neandertal ancestry excess a threshold of 0.90 while the second, la, estimates averages the probability of Neandertal ancestry.

To compute ta, at each SNP j within the locus of interest, we counted the fraction of individuals that fall on a Neandertal haplotype (which in turn is defined as a run of SNPs assigned marginal probability of at least 0.90). We interpolate this statistic linearly between SNPs and compute the average value of this function over the locus. A non-zero value indicates that at least one of the samples shows strong evidence of introgression at this locus. For our second statistic la, at each SNP j within the locus, we compute the average of the probability of the Neandertal ancestry across all haplotypes, $la(j) = \frac{\sum_{s=1}^{m} \gamma_{s,j}}{m}$. We compute the average value of this function over the locus after linear interpolation.

Locus	Coordinates	EUR		ASN	
		ta (%)	la (%)	ta (%)	la (%)
HLA class I	6:29,691,240-31,324,934	0.43	1.57	0.43	1.57
STAT2	12:56,737,172-56,750,354	6.60	7.23	3.90	4.91
OAS	12:113,344,844-113,448,288	26.45	30.63	26.45	30.63
SLC16A11	17:6,944,997-6,946,903	1.70	5.40	8.00	13.93
DMD	X:31,139,949-33,229,428	0.76	4.84	0.36	4.77

Table SI 11.1: Neandertal ancestry predicted by CRF at loci previously found to be introgressed.

Table SI 11.1 estimates ta and la in European and East Asian populations for five loci that have been reported to contain Neandertal introgression in previous studies. We see from Table SI 11.1 that all the loci show evidence for introgression according to the predictions of the CRF ($ta > 0$).

We do observe some differences that might reflect the limitations of the CRF. Abi-Rached et al. (2011) estimate the contribution of archaic ancestry at HLA-A to be $> 50\%$ in Europeans and $> 70\%$ in Asians. The CRF predicts Neandertal ancestry at this locus to be about 1.5% which is close to the background. One reason for this difference could be due to the action of balancing selection at this locus. As a result, alleles are likely to be shared between African, non-Africans and Neandertals even if introgression occurred. Thus, the CRF might be expected to have a high false negative rate at such loci. Mendez et al. (2012) show, using diagnostic SNPs that tag the introgressed haplotype, that the introgressed haplotype at STAT2 has a frequency in Eurasian populations ranging from $2 - 9\%$. Mendez et al. (2012) also show that there are two variants of the introgressed haplotype – a long and a short variant. They provide evidence for positive selection on the long variant in Melanesians. Mendez et al. (2013) show that the average allele frequency of 4 diagnostic SNPs that tag the introgressed Neandertal haplotype in the European HGDP populations is about 30%. We observe similar frequencies at this locus in European and East Asian populations. At both these loci, there is evidence of introgression from Denisovans as well (Mendez et al., 2013, 2012). Similar maps of Denisovan introgression would be useful in characterizing these events on a genomewide scale.

A Maximum likelihood estimate of the drift since Neandertal gene flow

We consider the problem of computing the parameters that maximize the log likelihood $\mathcal{L}(c, \alpha, \tau)$ defined in 7 where $\alpha \in [\alpha_l, \alpha_u], \tau \in [\tau_l, \tau_u]$. $\mathcal{L}$ is not convex. To maximize $\mathcal{L}$, we use a grid-search followed by a refinement approach.

To maximize $\mathcal{L}$, we consider a grid of width $(\alpha_{inc}, \tau_{inc})$ over the interval $\mathcal{I} = [\alpha_l, \alpha_u] \times [\tau_l, \tau_u]$. At each point (α, τ) in the grid, we compute the profile log likelihood $\mathcal{L}_p(\alpha, \tau) = max_c \mathcal{L}(c, \alpha, \tau)$. For fixed (α, τ), $\mathcal{L}$ is strongly convex in c. The profile log likelihood can be computed analytically for each (α, τ). If the maximum of the profile log likelihood is attained at (α_1, τ_1), we consider a new interval $\mathcal{I}_1 = [{\alpha_l}^1, {\alpha_u}^1] \times [{\tau_l}^1, {\tau_u}^1]$. Here ${\alpha_l}^1 = max(\alpha_1 - \alpha_{inc}, \alpha_l)$, ${\alpha_u}^1 = min(\alpha_1 + \alpha_{inc}, \alpha_u)]$,${\tau_l}^1 = max(\tau_1 - \tau_{inc}, \tau_l)$,${\tau_u}^1 = min(\tau_1 + \tau_{inc}, \tau_u)$. We maximize $\mathcal{L}_p$ over $\mathcal{I}_1$ using a Nelder-Mead Simplex algorithm (Nelder and Mead, 1965; Galassi et al) to obtain the MLE $(\hat{\alpha}, \hat{\tau})$. We set $t_{inc} = 0.01, f_{inc} = 0.005$, $\tau_l = 0.01, \tau_u = 0.30$. We set α_l, α_u based on prior knowledge. In simulations, we set $\alpha_l = 0.01, \alpha_u = 0.04$. For application to empirical data, we set bounds on $\alpha_l = 0.00, \alpha_u = 0.04$.

References

URL http://www.broadinstitute.org/mpg/cmsviewer/download/cms_localized_regions_062712.txt.

URL http://www.stat.ucla.edu/~handcock/RelDist.

URL http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz.

URL http://www.maths.bris.ac.uk/~wavethresh/.

L. Abi-Rached, M. J. Jobin, S. Kulkarni, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, 334(6052):89–94, Oct 2011.

M. Abramowitz and I. Stegun. *Handbook of Matehmatical Functions*.

Y. Baran, B. Pasaniuc, S. Sankararaman, et al. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 2012.

J. C. Barrett, S. Hansoul, D. L. Nicolae, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, 40(8):955–962, Aug 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

S. Bercovici, J. Rodriguez, M. Elmore, and S. Batzoglou. Ancestry inference in complex admixtures via variable-length markov chain linkage models. 7262:12–28, 2012.

S. Bochkanov and V. Bystritsky. Alglib (www.alglib.net).

A. Brisbin, K. Bryc, J. Byrnes, et al. Pcadmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, 84(4), 2012.

S. Browning and B. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

K. Bryc, C. Velez, T. Karafet, et al. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*, 2010.

F. Busing, E. Meijer, and R. Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9:3–8, 1999.

R. H. Byrd, J. Nocedal, R. B. Schnabel, R. H. B. J. Nocedal, and R. B. Representations of quasi-newton matrices and their use in limited memory methods, 1994.

B. Charlesworth, M. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

D. Charlesworth, B. Charlesworth, and M. Morgan. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–1632, 1995.

S. A. Chung, K. E. Taylor, R. R. Graham, et al. Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet.*, 7(3):e1001323, Mar 2011.

I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.

H. Furberg, Y. Kim, J. Dackor, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.*, 42(5):441–447, May 2010.

M. Galassi et al. *GNU Scientific Library Reference Manual*. 3 edition.

C. Gini. *Italian: Variabilità e mutabilità" (Variability and Mutability'*. 1912.

S. Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

R. E. Green, J. Krause, A. W. Briggs, et al. A draft sequence of the neandertal genome. *Science*, 328(5979):710–722, 2010.

S. R. Grossman, I. Shlyakhter, I. Shylakhter, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–886, Feb 2010.

R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, 5(10):e1000695, Oct 2009.

M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, and J. D. Wall. Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U.S.A.*, 108(37):15123–15128, Sep 2011.

K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.*, 9(6):e1003521, Jun 2013.

M. He, M. C. Cornelis, P. Kraft, et al. Genome-wide association study identifies variants at the IL18-BCO2 locus associated with interleukin-18 levels. *Arterioscler. Thromb. Vasc. Biol.*, 30(4): 885–890, Apr 2010.

G. Hellenthal and M. Stephens. mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4):520–521, 2007.

G. Hellenthal, A. Auton, and D. Falush. Inferring human colonization history using a copying model. *PLoS Genet*, 4(5):e1000078, 05 2008.

A. G. Hinch, A. Tandon, N. Patterson, et al. The landscape of recombination in African Americans. *Nature*, 476:170–175, Aug 2011.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

A. S. Hinrichs, D. Karolchik, R. Baertsch, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, 34(Database issue):D590–598, Jan 2006.

R. R. Hudson. Generating samples under a wright-fisher neutral model. *Bioinformatics*, 18:337–338, 2002.

N. A. Johnson, M. A. Coram, M. D. Shriver, et al. Ancestral components of admixed genomes in a mexican cohort. *PLoS Genet*, 7(12):e1002410, 12 2011.

A. Keinan, J. C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.*, 39 (10):1251–1255, Oct 2007.

E. E. Kenny, I. Pe'er, A. Karban, et al. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet.*, 8(3):e1002559, 2012.

M. Kimura. Solution of a process of random genetic drift with a continous model. *Proc. Natl. Acad. Sci. U.S.A.*, 41(3):144–150, Mar 1955.

A. Kong, G. Thorleifsson, D. F. Gudbjartsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467:1099–1103, Oct 2010.

H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

J. Lachance, B. Vernot, C. C. Elbers, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell*, 150(3):457–469, August 2012.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Y. H. Lee, S. C. Bae, S. J. Choi, J. D. Ji, and G. G. Song. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. *Mol. Biol. Rep.*, 39(12):10627–10635, Dec 2012.

H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

H. Li, B. Handsaker, A. Wysoker, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th international conference on Machine learning*, pages 584–591. ACM, 2008.

S. Macgregor, A. W. Hewitt, P. G. Hysi, et al. Genome-wide association identifies ATOH7 as a major gene determining human optic disc size. *Hum. Mol. Genet.*, 19(13):2716–2724, Jul 2010.

B. Maples, S. Gravel, E. Kenny, and C. Bustamante. Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, 93:278–288, 2013.

G. McVicker, D. Gordon, C. Davis, and P. Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5(5):e1000471, 05 2009.

G. Mells et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nature Genetics*, 43(4):329–32, 2011.

F. L. Mendez, J. C. Watkins, and M. F. Hammer. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.*, 91(2):265–274, Aug 2012.

F. L. Mendez, J. C. Watkins, and M. F. Hammer. Neandertal Origin of Genetic Variation at the Cluster of OAS Immunity Genes. *Mol. Biol. Evol.*, 30(4):798–801, Apr 2013.

M. Meyer, M. Kircher, M.-T. Gansauge, et al. A high-coverage genome sequence from an archaic denisovan individual. *Science*, 2012.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7 (4):308–313, 1965.

A. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, 12(6):443–451, Jun 2011.

B. Paten, J. Herrero, S. Fitzgerald, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18(11):1829–1843, Nov 2008.

D. Percival and A. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2005.

A. L. Price, A. Tandon, N. Patterson, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 06 2009.

K. Prüfer, F. Racimo, N. Patterson, et al. The complete genome sequence of a neandertal from the altai mountains. *submitted*, 2013.

D. Reich, R. E. Green, M. Kircher, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468:1053–1060, Dec 2010.

D. Reich, N. Patterson, D. Campbell, et al. Reconstructing Native American population history. *Nature*, 488(7411):370–374, Aug 2012.

S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *American Journal of Human Genetics*, 8(2):290–303, 2008.

S. Sankararaman, N. Patterson, H. Li, S. Pääbo, and D. Reich. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.*, 8(10):e1002947, 2012.

S. A. Sawyer and D. L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132 (4):1161–76, 1992.

A. Scally, J. Y. Dutheil, L. W. Hillier, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175, 2012.

P. Skoglund and M. Jakobsson. Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. U.S.A.*, 108(45):18301–18306, Nov 2011.

K.-A. Sohn and E. P. Xing. Spectrum: joint Bayesian inference of population structure and recombination events. *Bioinformatics*, 23:479–489, 2007.

A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4):676–682, 2008.

C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 2011. To appear.

H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79:1–12, 2006.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

The SIGMA Type 2 Diabetes Consortium. Sequence variants in SLC16A11 are a comon risk factor for type 2 diabetes in mexico. *Under review.*

J. D. Wall. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154(3):1271–1279, Mar 2000.

J. D. Wall, K. E. Lohmueller, and V. Plagnol. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution*, 2009.

J. D. Wall, M. A. Yang, F. Jay, et al. Higher Levels of Neanderthal Ancestry in East Asians Than in Europeans. *Genetics*, Feb 2013.

S. Williamson and M. E. Orive. The genealogy of a sequence subject to purifying selection at multiple sites. *Molecular biology and evolution*, 19(8):1376–1384, 2002.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*, 2010.

M. A. Yang, A. S. Malaspinas, E. Y. Durand, and M. Slatkin. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol.*, 29(10):2987–2995, Oct 2012.

V. Yotova, J. F. Lefebvre, C. Moreau, et al. An X-linked haplotype of Neandertal origin is present among all non-African populations. *Mol. Biol. Evol.*, 28(7):1957–1962, Jul 2011.