SUPPLEMENTARY INFORMATION - 2

Inference procedure

Ivan Juric, Simon Aeschbacher, Graham Coop

1 Introduction

In our S1 Text we described how the present-day frequency of Neanderthal alleles depends on the selection coefficient, s, recombination rate, and the initial frequency of Neanderthal allele, p_0. Here, we introduce the last model parameter, μ, the average probability that a Neanderthal allele at exonic site is deleterious. We then discuss the details of our inference procedure, and expand on our results.

2 Theory

2.1 Incorporating the probability that an exonic site contains a deleterious Neanderthal allele (μ)

Some loci are more likely to harbor deleterious alleles. For example, mutations at exonic sites are more likely to affect fitness than mutations in introns or in nongenic regions. In all our models we assume that the Neanderthal allele at exonic sites can be deleterious, while all other Neanderthal alleles are neutral. To that end, we introduce μ, the average probability that an exonic Neanderthal allele is selected against in modern humans. When μ is small, it is also approximately equal to the average density of deleterious Neanderthal exonic sites.

Consider now a neutral locus k on an autosomal chromosome surrounded by

i exonic sites in some predefined window. Let the recombinational distance of the closest exonic site (regardless of whether it is left or right of neutral locus) be labeled r_1, the next one r_2 and so on such that the exonic site farthest from the neutral locus is assigned distance r_i. If deleterious exonic sites are rare, two such sites will be so far apart that a neutral site k will be affected almost exclusively by the closest deleterious site. Therefore, we can approximate the expected frequency of Neanderthal allele at a neutral locus by considering only the effect of the closest selected exonic site. In this case the probability that the j^{th}, $1 \leq j \leq i$, closest exonic selected site is deleterious is $\mu(1-\mu)^{j-1}$ and $p_{t,k}$, the frequency of Neanderthal allele at a neutral locus linked to j^{th} exonic site, is equal to $x_0 f(s, r_j, t) + y_0$ (equation 6 in S1).

We can calculate the expected frequency of Neanderthal allele at a neutral locus k, surrounded by i exonic sites, by calculating p_t for all cases when one or none of i exonic sites in a window is under selection

$$\begin{aligned} E[p_{k,t}] &= \sum_{j=1}^{i} \mu(1-\mu)^{j-1}(x_0 f(s, r_j, t) + y_0) + (1-\mu)^i(x_0 + y_0) \\ &= x_0(\mu \sum_{j=1}^{i} (1-\mu)^{j-1} f(s, r_j, t) + (1-\mu)^i) + y_0 \end{aligned} \tag{1}$$

The term $(1-\mu)^i(x_0 + y_0)$ in the top line of equation (1) accounts for the case when none of i exonic sites in our window are selected. To obtain the expression for the expected Neanderthal allele frequency on X chromosome, we replace x_0, y_0 and $f(s, r_j, t)$ in equation (1) by terms corresponding to X chromosome (see the right-hand side of equation 12 in S1). In practice we find μ to be small, which allows us estimate $E[p_t, k]$ by summing over exons rather than all exonic sites. This approximation substantially increases computational efficiency, and leads to equation (8), described later in this supplement.

2.2 Inference procedure

2.2.1 Residual sum of squares (RSS)

Our inference method relies on finding the parameters that minimize the residual sum of squared differences (RSS) between the observed p_n from Sankararaman et al. [2012] and the expected $E[p_t]$ Neanderthal allele frequencies across all SNPs in a present-day human sample. Our RSS is given by

$$\text{RSS} = \sum_{k=1}^{n_l} (p_{k,n} - E[p_{k,t}])^2 = \sum_{k=1}^{n_l} (p_{k,n} - x_0 g_k(\mu, s, \boldsymbol{r}, t) - y_0)^2, \tag{2}$$

where the sum is over all n_l SNPs in the autosomal genome (or on the X). We can rewrite the equation (2) as:

$$\text{RSS} = \sum_{k=1}^{n_l} p_{k,n}^2 - 2x_0 \sum_{k=1}^{n_l} p_{k,n} g_k - 2y_0 \sum_{k=1}^{n_l} p_{k,n} + x_0^2 \sum_{k=1}^{n_l} g_k^2 + 2x_0 y_0 \sum_{k=1}^{n_l} g_k + y_0^2 n_l, \tag{3}$$

where, for clarity, we write g_k instead of $g_k(\mu, s, \boldsymbol{r}, t)$. When $y_0 = 0$, i.e. the deleterious allele is fixed in Neanderthals, equation (3) simplifies to:

$$\text{RSS} = \sum_{k=1}^{n_l} p_{k,n}^2 - 2p_0 \sum_{k=1}^{n_l} g_k p_{k,n} + p_0^2 \sum_{k=1}^{n_l} g_k^2 \tag{4}$$

2.2.2 Minimizing RSS

For a given μ and s all of the summations in equation (3) are constants, and the minimum RSS depends only on x_0 and y_0. This function, which we will refer to $\text{RSS}(x_0, y_0)$, has a minimum if $D(x_c, y_c) > 0$ and $\text{RSS}_{xx}(x_c, y_c) > 0$, where:

$$D(x_c, y_c) = \text{RSS}_{xx}(x_c, y_c)\text{RSS}_{yy}(x_c, y_c) - \left[\text{RSS}_{xy}(x_c, y_c)\right]^2, \tag{5}$$

and $\text{RSS}_{xx}(x_c, y_c)$, $\text{RSS}_{yy}(x_c, y_c)$ and $\text{RSS}_{xy}(x_c, y_c)$ are second order partial derivative of RSS with respect to x_0, y_0 and the cross partial derivative with respect to x_0 and y_0, all evaluated at the critical points x_c, y_c. We obtain the critical points by solving a system of first order derivatives: $\text{RSS}_x(x_0, y_0) = 0$

and $\mathrm{RSS}_y(x_0, y_0) = 0$. We find:

$$x_c = \frac{\sum p_{k,n} \sum g_k - n_l \sum p_{k,n} g_k}{\left[\sum g_k\right]^2 - n_l \sum g_k^2} \tag{6a}$$

$$y_c = \frac{\sum p_{k,n} g_k \sum g_k - \sum p_{k,n} \sum g_k^2}{\left[\sum g_k\right]^2 - n_l \sum g_k^2} \tag{6b}$$

$$\mathrm{RSS}_{XX}(x_c, y_c) = 2 \sum g_k^2 \tag{6c}$$

$$D = 4 \left(n_l \sum g_k^2 - \left[\sum g_k\right]^2 \right) \tag{6d}$$

where we have dropped index over the summations for clarity. By the Cauchy-Schwarz inequality, D is always positive. The $\mathrm{RSS}_{xx}(x_c, y_c)$ is also positive since $g_k > 0$. Therefore x_c and y_c minimize RSS for a given μ and s.

Notice that mathematically, x_c and y_c can be less than zero. A negative value of x_c will occur if $\sum p_{k,n}(\overline{g} - g_k) > 0$ while $y_c < 0$ if $\sum p_{k,n}(g_k \overline{g} - \overline{g^2}) > 0$, where $\overline{g}$ is the mean of function g, $\overline{g} = (1/n_l) \sum g_k$ and $\overline{g^2}$ is the mean of g_k^2. In reality, minimal initial frequency of N_1S_1 (x_0) and N_1S_2 (y_0) haplotypes cannot be lower than zero. Given that constraint and a value of x_c or y_c less than zero, RSS is minimized when the "negative" haplotype frequency is equal to zero. In that case, when the initial frequency of haplotype carrying deleterious allele is zero ($x_0 = 0$), the expected initial frequency of the other,y, haplotype is the same as the average present day frequency of neutral Neanderthal allele N_1 because the population is forever monomorphic at selected locus S. On the other hand, when $y_c = 0$, equation (3) turns into equation (4), which can then be minimized with respect to x_0. With caveat involving negative values explained, we see that x_c and y_c are the estimated Neanderthal haplotype frequencies which minimize the $\mathrm{RSS}(x_0, y_0)$ for a given s and μ.

For a given single locus model described in S1 Text, we evaluate the RSS for a particular value for μ and s, denoted by μ_i and s_j as follows:

1. Calculate x_c and y_c using equation (6) given μ_i and s_j.

2. If $x_c < 0$, set x_c to zero and y_c to mean observed Neanderthal frequency, $\overline{p_n}$.

3. If $y_c < 0$, set y_0 to zero and x_c to $\sum p_{k,n} g_k / \sum g_k^2$.

4. Calculate RSS using equation (3) by replacing x_0 and y_0 with x_c and y_c.

We repeat steps 1 -4 for all combinations of μ and s. We find the smallest value of RSS over all combinations of tested values of μ and s.

3 Technical implementation

Recently, Sankararaman et al. [2012] estimated p_n, the frequency of Neanderthal ancestry in modern-day Europeans (EUR) and East Asians (ASN) at numerous SNPs in the genome. We downloaded those estimates, as well as physical and genetic position of SNPs from the Reich Lab website. The genetic map resolution is 1×10^{-3} cM, so if two loci are closer than that distance, they are assigned the same position. We downloaded a list of exons from UCSC Genome Bioinformatics browser assembly (hg19) to match the files containing p_n.

For genes with alternative splicing, we collapsed all overlapping exons to create exonic regions, each of which starts at the beginning of the leftmost overlapping exon and ends at the end of the right-most overlapping exon. For genes without alternative splicing, our exonic regions are equivalent to exons. In the rest of text, we will refer to exonic regions as exons.

For each site, we consider only exons whose midpoint is within 1 cM window from the focal neutral locus (0.5 cM on each side). We give the justification for this in the main text. Increasing this window size by a factor of 10 did not change the estimates substantially for our single locus equilibrium model, but it increased the computational time substantially. We also tried estimated parameters by fitting a RSS model to the mean observed and expected allele frequency in 0.1 cM non-overlapping blocks. We found that parameters estimated from

such model agree well with the results based on considering each SNP independently, this suggests that our choice to calculate our RSS on a per-SNP level has not affected our analysis.

We use linear interpolation to determine the genetic map position of each exon midpoint. For an exon starting at physical position x_1 and ending at position x_2, we first find the midpoint x_m for which we calculate its position on the genetic map as:

$$r_m = r_1 + (r_2 - r_1)(x_m - y_1) \tag{7}$$

where r_1 and r_2 are the genetic map position of the closest SNP to the left and right of x_m and y_1 is the physical position of the closest left SNP. For exons that are positioned left of first SNP, we assumed that the recombination rate per base pair between such exons and first SNP is the same as between first and second SNP. Similarly, for exons right of the last SNP, we assumed that recombination rate per base pair is the same as between the last and next to last SNPs. In the cases when first SNP was assigned genetic position 0 but there were exons left of that SNP, we removed the SNP and exons were assigned position 0.

Lastly, if μ is small, the probability that an exon of length $l$bp contains selected site is approximately μl. We use this approximation to speed up the calculation of p_t. Most exons are short and $\mu l \ll 1$. However, if exons are so long that $\mu l \approx 1$, we divide exons repeatedly in half until this condition holds. In the end, for a given SNP k, we approximate equation (1) by

$$E[p_{t,k}] = x_0 \left(\sum_{i=1}^{i_k} \mu l_i \left[\prod_{j=1}^{i-1} (1 - \mu l_j) \right] f(s, r_i, t) + \prod_{j=1}^{i_k} (1 - \mu l_j) \right) + y_0, \tag{8}$$

where i_k is the number of exons whose midpoints are within a window surrounding SNP k. When $y_0 = 0$, i.e. the deleterious allele is fixed in Neanderthals, then $x_0 = p_0$ and

$$E[p_{t,k}] = p_0 \left(\sum_{i=1}^{i_k} \mu l_i \left[\prod_{j=1}^{i-1} (1 - \mu l_j) \right] f(s, r_i, t) + \prod_{j=1}^{i_k} (1 - \mu l_j) \right), \tag{9}$$

which can be written as $E[p_{k,t}] = p_0 g_k(\mu, s, \boldsymbol{r}, t)$. This simplified version of the equation (8) is presented in the main text.

To calculate the RSS for the multiple locus model at each of the 676 (26^2) combinations of μ and s we drew 30 replicates of n selected sites for each chromosome and distributed those sites randomly across the exonic sites. The number of selected sites per chromosome, n, comes from the binomial distribution with parameters μ and n_{tot}, where n_{tot} is the total number of exonic sites on the chromosome. Then, for each SNP, we calculated p_t using the equation (16) from S1, and take the mean of this across our 30 replicates. This mean of p_t at each SNP is then used in the calculation of the RSS.

4 Results

4.1 Model fit

Tables S2.1 and S2.2 contain the point estimates and 95% block bootstrap confidence intervals for parameters that minimize RSS under different models. Our bootstrap method is explained in the methods section of the main text. Figures S2.1 – S2.5 show the RSS surfaces for μ and s for different models, while figure S2.6 shows the RSS surfaces for the initial frequency of Neanderthal allele on X chromosome $p_{0,X}$.

In figures S2.7 – S2.9 we show the fit between the average observed frequency of Neanderthal alleles, binned by gene density per map unit, and the allele frequency predicted by our model. Each plot is created by first splitting the genome into segments of constant size (in cM), then counting the number of exonic sites in each segment and lastly binning segments into bins of equal size. Table S2.3 gives the Pearson correlation coefficient between observed and model predicted average Neanderthal allele frequency across all bins for a range of bin sizes. For each bin we calculate the average observed frequency of Neanderthal alleles and the frequency predicted by our model using parameters from table S2.1.

Population	Scenario	p_0	s	μ
EUR	Single Locus, $t = \infty$	0.0315	1.02×10^{-4}	7.8×10^{-5}
EUR	Multiple Loci, $t = \infty$	0.0312	2.5×10^{-5}	1.2×10^{-4}
EUR	Single Locus, $t = 2000$	0.0338	4.12×10^{-4}	8.1×10^{-5}
EUR	Single Locus, $t = 2000$, X chr	0.0292	9.6×10^{-4}	8.1×10^{-5}
ASN	Single Locus, $t = \infty$	0.0349	8.8×10^{-5}	6.8×10^{-5}
ASN	Multiple Loci $t = \infty$	0.0350	3.7×10^{-5}	1.0×10^{-4}
ASN	Single Locus, $t = 2000$	0.0360	3.52×10^{-4}	6.9×10^{-5}
ASN	Single Locus, $t = 2000$, X chr	0.0298	1.6×10^{-4}	6.8×10^{-4}

Table S2.1: Minimum RSS parameters for μ, s and p_0 for different models described in S1 Text. Figure 1 in the main text shows an example of $E[p_t]$ for single locus model, $t = 2000$, for part of chromosome 1.

Population	Scenario	p_0	s	μ
EUR	Single Locus, $t = \infty$	$[3.00, 3.30] \times 10^{-2}$	$[0.8, 1.4] \times 10^{-4}$	$[7.8, 8.1] \times 10^{-5}$
EUR	Single Locus, $t = 2000$	$[3.22, 3.52] \times 10^{-2}$	$[3.4, 5.2] \times 10^{-4}$	$[0.41, 1.2] \times 10^{-4}$
EUR	Single Locus, $t = 2000$, X chr	$[2.32, 3.53] \times 10^{-2}$	$[0.64, 2.08] \times 10^{-3}$	$[0.41, 1.6] \times 10^{-4}$
ASN	Single Locus, $t = \infty$	$[3.35, 3.66] \times 10^{-2}$	$[0.40, 1.6] \times 10^{-4}$	$[4.1, 8.1] \times 10^{-5}$
ASN	Single Locus, $t = 2000$	$[3.45, 3.86] \times 10^{-2}$	$[2.6, 5.4] \times 10^{-4}$	$[0.41, 1.6] \times 10^{-4}$
ASN	Single Locus, t = 2000, X chr	$[2.36, 3.9] \times 10^{-2}$	$[0, 4] \times 10^{-3}$	$[0.001, 1] \times 10^{-3}$

Table S2.2: The 95% bootstrap confidence intervals for μ, s, and p_0 for different models.

4.2 The X chromosome and sex-bias during admixture

In this subsection we describe how we estimated the sex-bias during admixture. We say that admixture is Neanderthal male-biased if more than 50% of introgressed alleles came from male Neanderthals. Conversely, if less than

Segment size	ρ_{EUR}	ρ_{ASN}
0.5 cM	0.871	0.881
1 cM	0.897	0.710
1.5 cM	0.887	0.546
2 cM	0.847	0.633

Table S2.3: Correlation between the estimated and the observed mean Neanderthal allele frequency for bins created using segments of different sizes.

50% introgressed alleles came from male Neanderthals, we say that admixture was Neanderthal female-biased. Otherwise, we say that admixture showed no sex-bias. Consider a single generation of matings between humans and Neanderthals. Let m_1 be the frequency of Neanderthal male, human female matings and let m_2 be the frequency of Neanderthal female, human male matings and let p_0 and $p_{0,x}$ be the initial frequency of Neanderthal autosomal and X linked alleles. Then

$$\begin{aligned} \frac{1}{2}m_1 + \frac{1}{2}m_2 &= p_0, \\ \frac{1}{3}m_1 + \frac{2}{3}m_2 &= p_{0,x}. \end{aligned} \tag{10}$$

The numbers multiplying m_1 and m_2 in the above are because an autosomal allele spends on average half of its time in males and half in females, while a X - linked allele spends 1/3s and 2/3s respectively. The solution of equation (10) is

$$\begin{aligned} m_1 &= 4p_0 - 3p_{0,x}, \\ m_2 &= 3p_{0,x} - 2p_0. \end{aligned} \tag{11}$$

Based on our estimates of p_0 and $p_{0,X}$ (Table S2.1) for EUR population we estimate $m_{1,EUR} = 0.0476$, $m_{2,EUR} = 0.02$, $m_{1,EUR}/m_{2,EUR} = 2.38$ and for ASN $m_{1,ASN} = 0.0546$, $m_{2,ASN} = 0.0174$, $m_{1,ASN}/m_{2,ASN} = 3.14$. We note that the CI intervals for these estimates are wide and include 1. However, in the main text (also see Figure 4 in the main text) we discuss that $\mu_X s_X$ and $p_{0,X}$ are confounded so it is possible that mating was sex-biased if the selection is a lot stronger on the X than autosomes. However, it seems likely that both selection and sex-biased mating may be in play in shaping X to autosome levels of admixture.

References

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

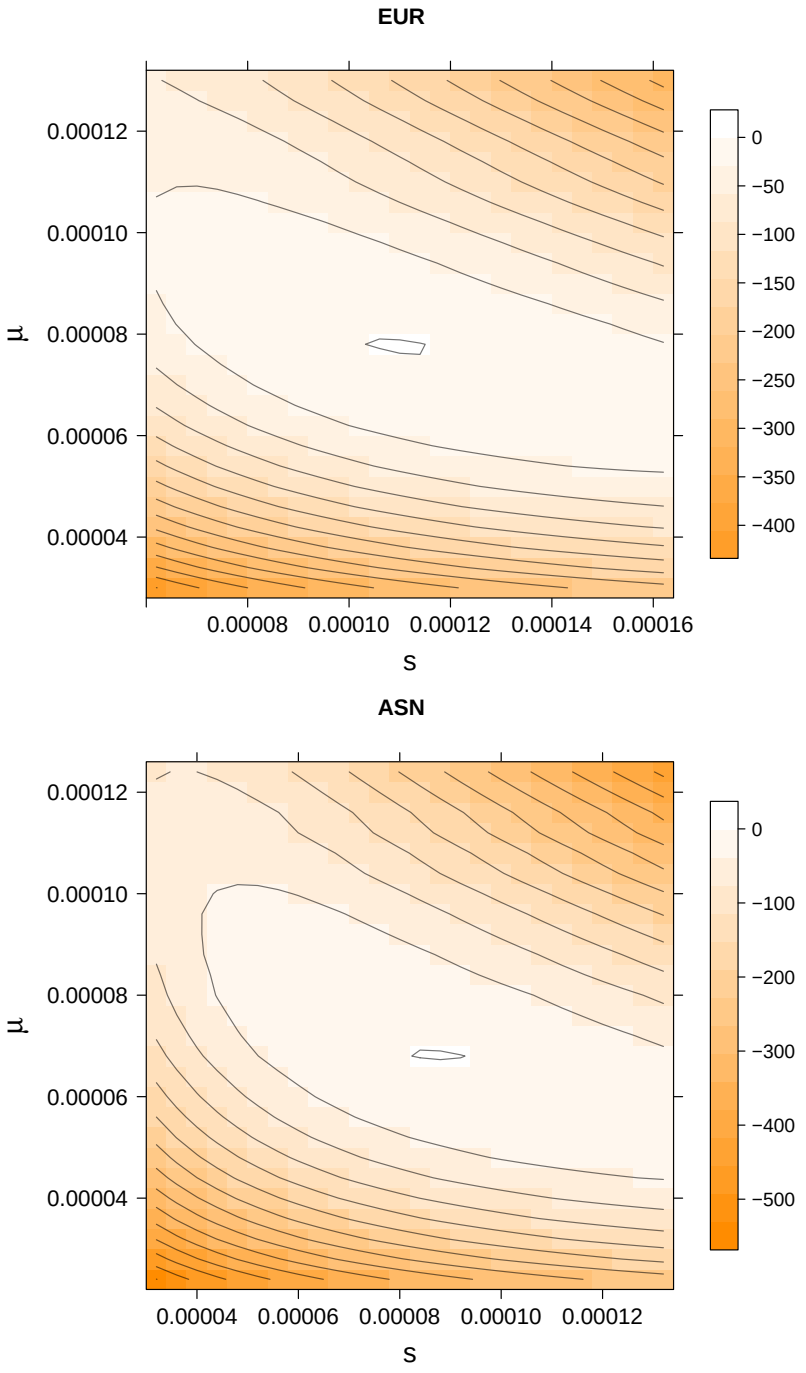

Figure S2.1: The scaled RSS surface ($\text{RSS}_{\min} - \text{RSS}$) for different s and μ values for EUR and ASN autosomal chromosomes for single locus equilibrium model ($t = \infty$). Each value of the RSS is minimized over p_0, making this a profile RSS surface. Regions shaded in orange represent parameter values of higher RSS.

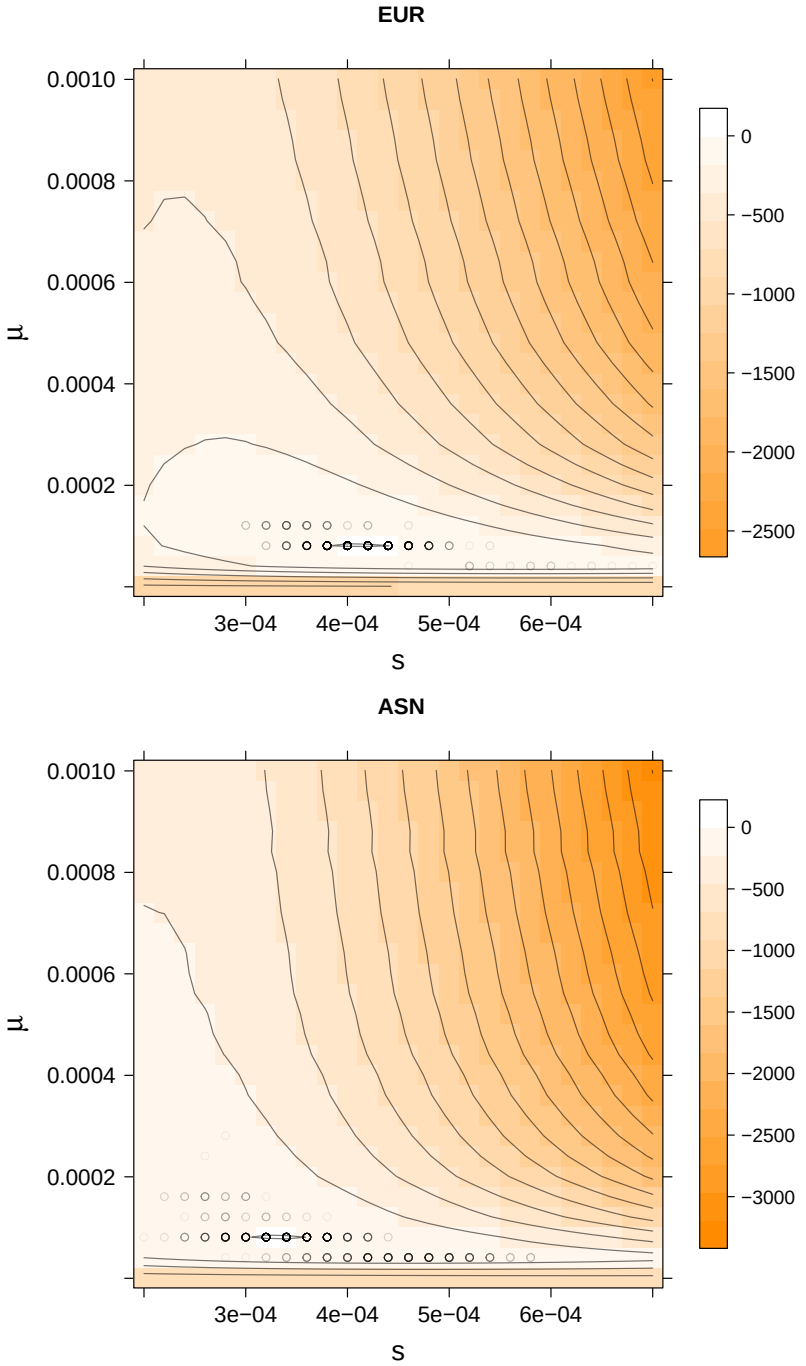

Figure S2.2: The scaled RSS surface ($\text{RSS}_{\min} - \text{RSS}$) for different s and μ values for EUR and ASN autosomal chromosomes for single locus model for $t = 2000$. Each value of the RSS is minimized over p_0, making this a profile RSS surface. Regions shaded in orange represent parameter values of higher RSS. Black circles show bootstrap results of 1000 block bootstrap reestimates, with darker circles corresponding to more common bootstrap estimates.

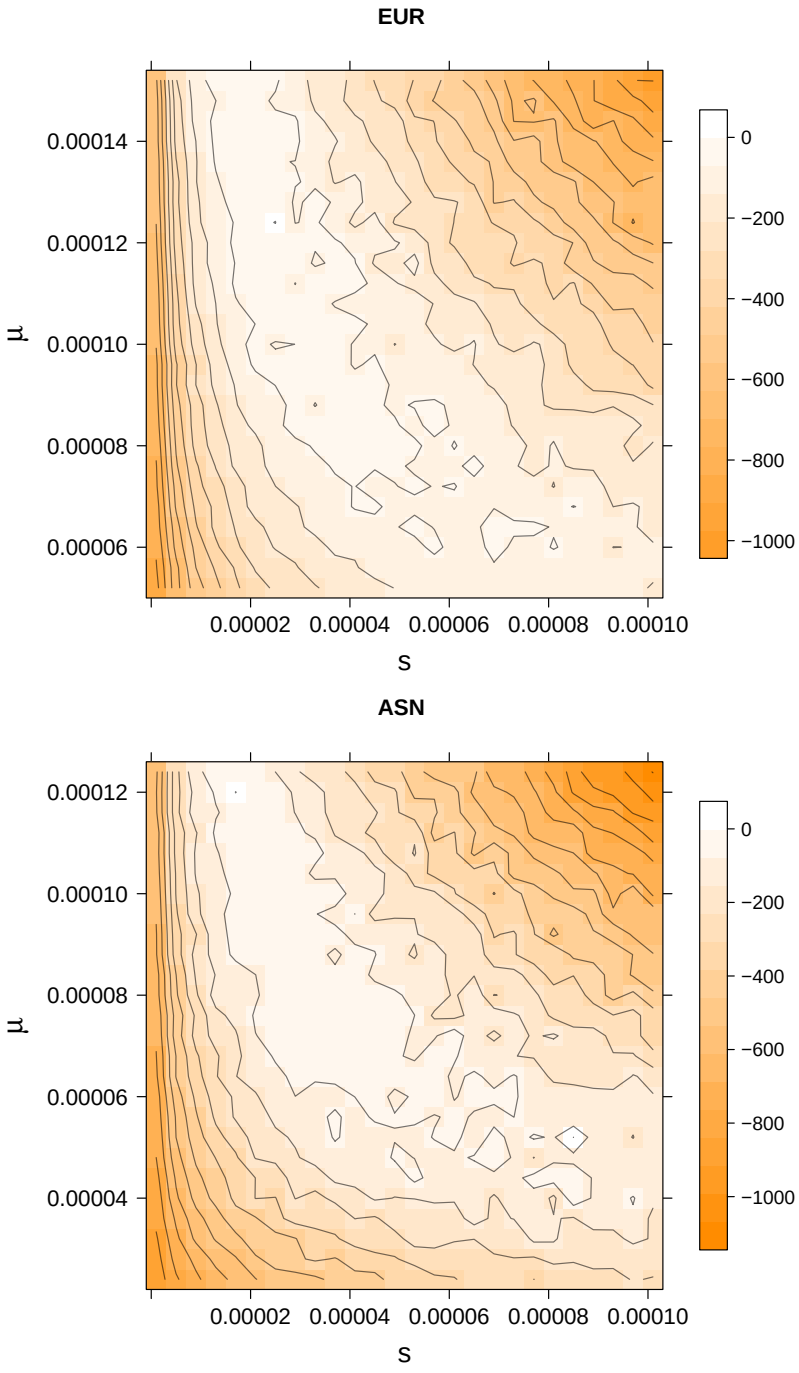

Figure S2.3: The scaled RSS surface ($\text{RSS}_{\min} - \text{RSS}$) for different s and μ values for EUR and ASN autosomal chromosomes for multi-locus equilibrium model ($t = \infty$). Each value of the RSS is minimized over p_0, making this a profile RSS surface. Regions shaded in orange represent parameter values of higher RSS.

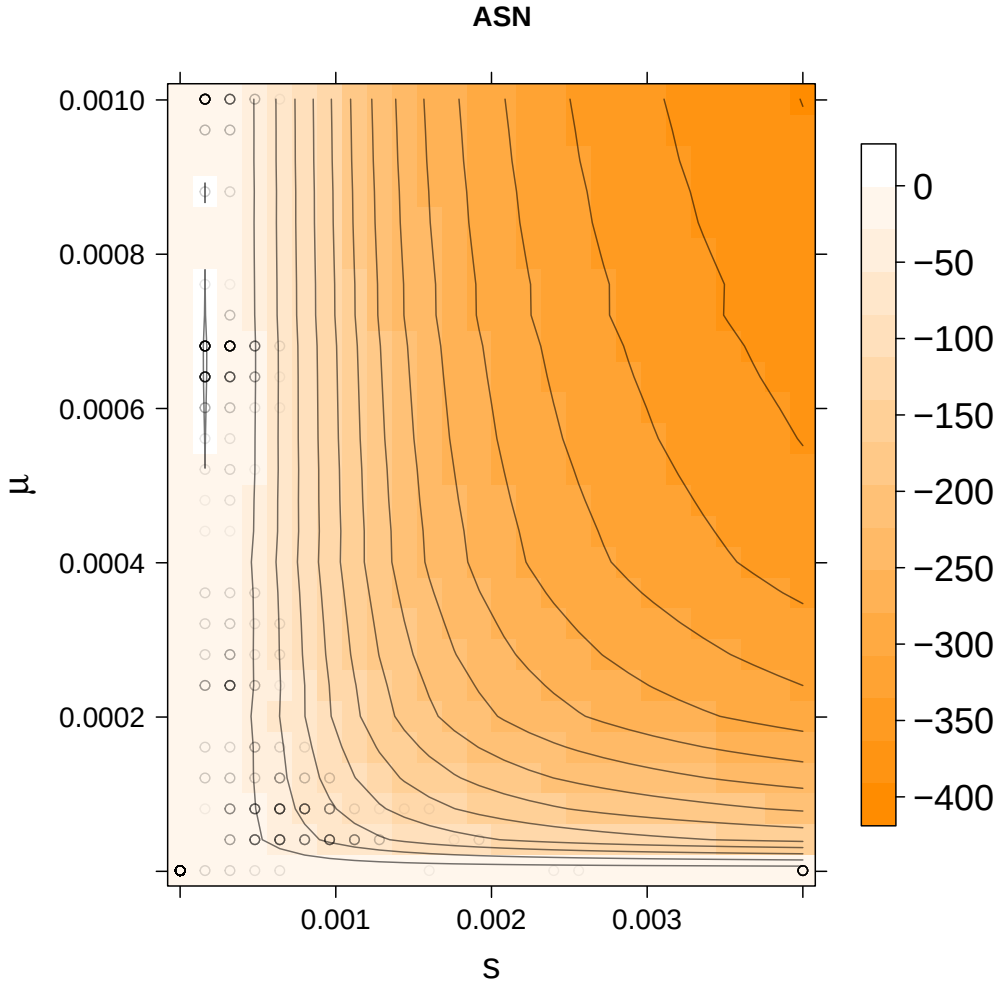

Figure S2.4: The scaled RSS surface ($\text{RSS}_{\min} - \text{RSS}$) for different s and μ values for ASN X chromosome for single locus model for $t = 2000$ and assuming equal strength of selection in males and females. Each value of the RSS is minimized over p_0, making this a profile RSS surface. Regions shaded in orange represent parameter values of higher RSS. Black circles show bootstrap results of 1000 block bootstrap reestimates, with darker circles corresponding to more common bootstrap estimates.

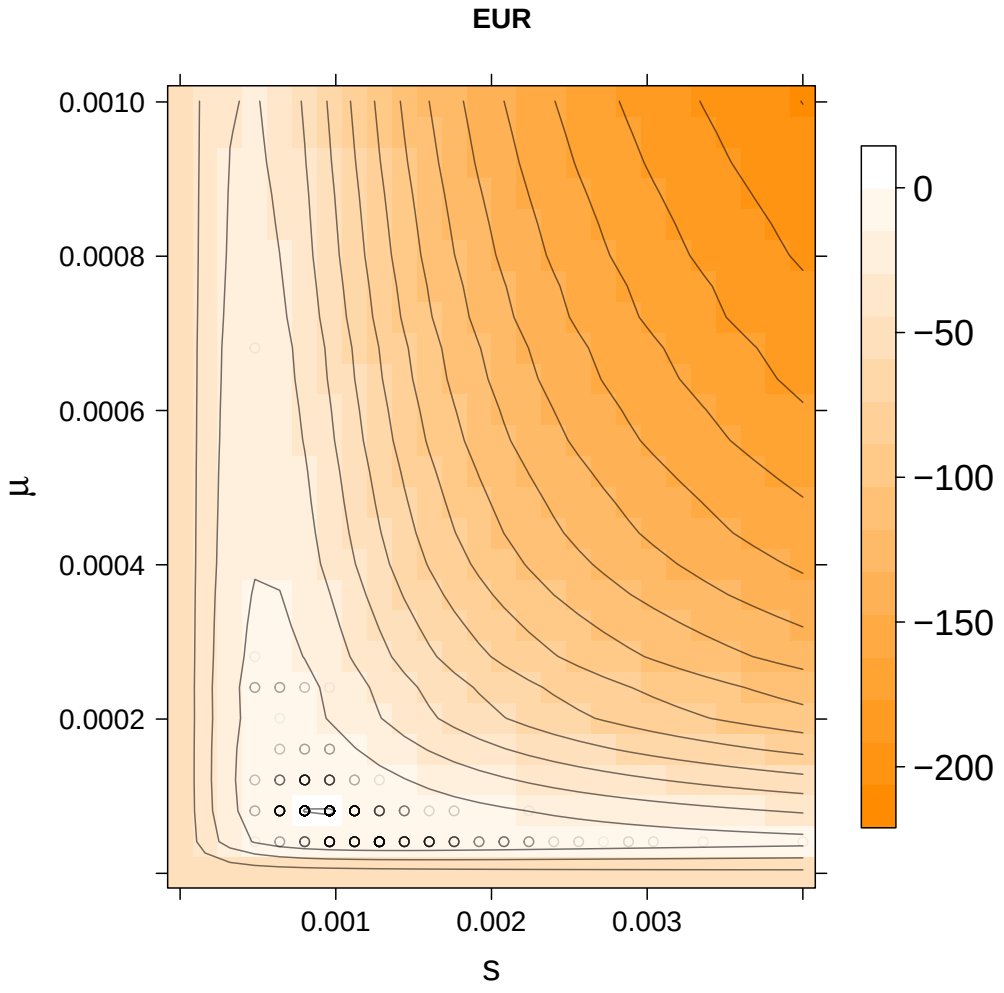

Figure S2.5: The scaled RSS surface ($\text{RSS}_{\min} - \text{RSS}$) for different s and μ values for EUR X chromosome for single locus model for $t = 2000$ and assuming equal strength of selection in males and females. Each value of the RSS is minimized over p_0, making this a profile RSS surface. Regions shaded in orange represent parameter values of higher RSS. Black circles show bootstrap results of 1000 block bootstrap reestimates, with darker circles corresponding to more common bootstrap estimates.

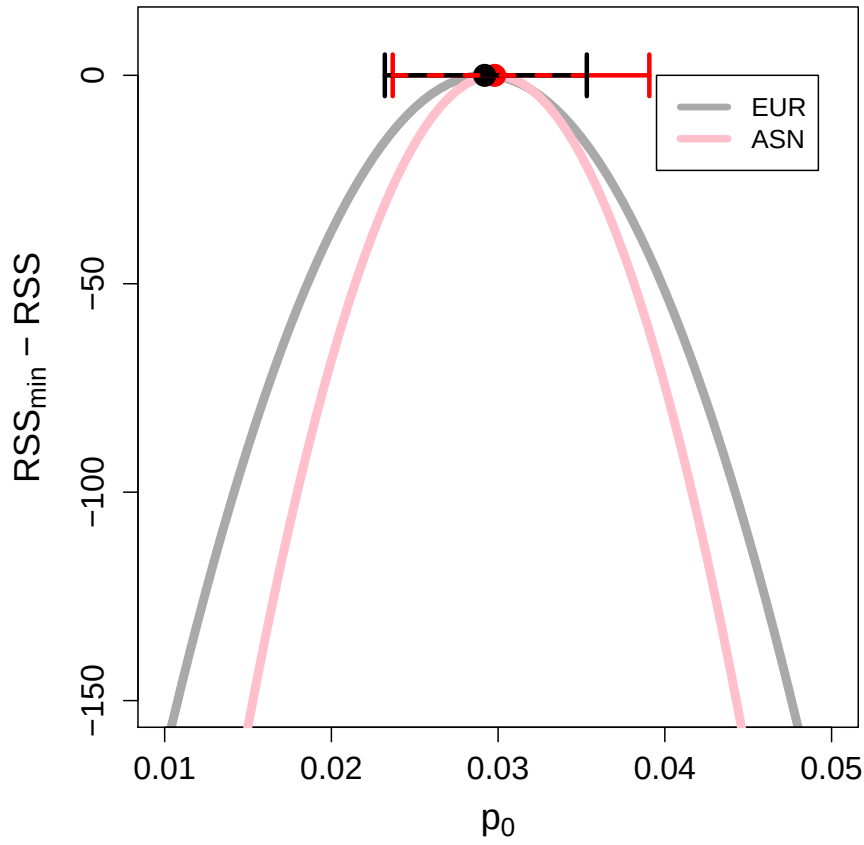

Figure S2.6: The scaled RSS surface ($\text{RSS}_{\text{min}} - \text{RSS}$) of X chromosomes as a function of the initial admixture proportion p_0. Results are shown for a model where only the nearest-neighboring exonic site under selection is considered, and for $t = 2000$ generations after Neanderthals split from EUR (grey) and ASN (pink) populations. Dots and horizontal lines show the value of p_0 that minimizes the RSS and the respective 95% block-bootstrap confidence intervals. Each value of the RSS is evaluated at the values of the selection coefficient (s) and exonic density of selection (μ) given in Table S2.1.

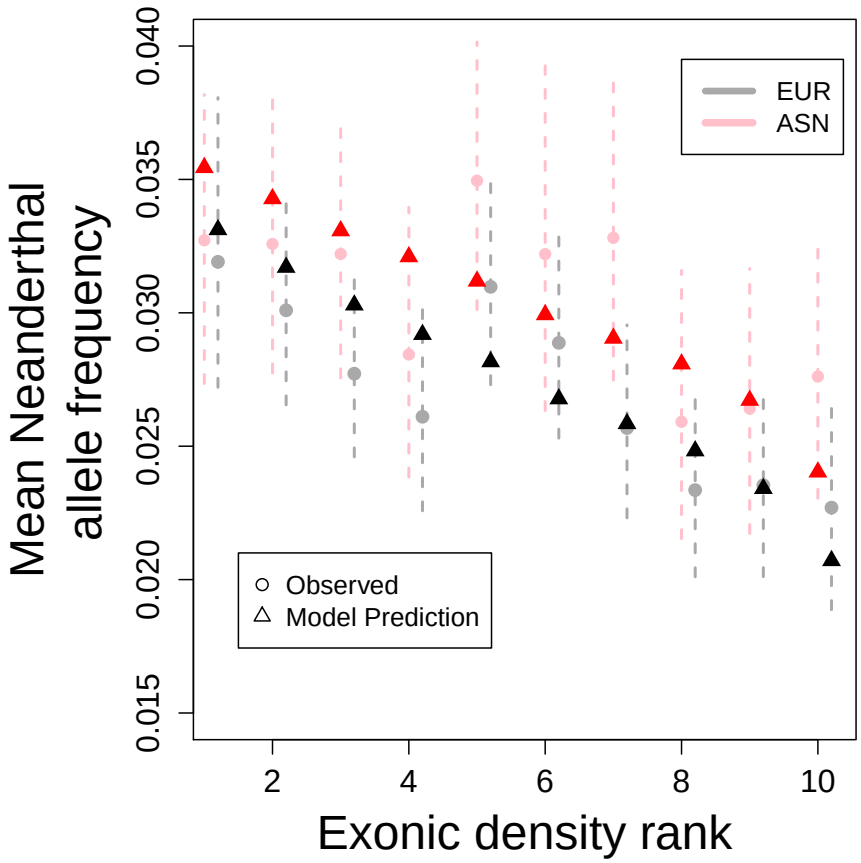

Figure S2.7: Fit between our estimates of p_t for bins of different exon density. Genomic regions with low exonic density (low exonic density rank) contain higher average Neanderthal allele frequency in both in Europeans (grey circle) and Asians (pink circle), a pattern recreated in our model. Dashed lines represent the 95% block bootstrap confidence intervals. The length of segments used to create the bins is 2 cM.

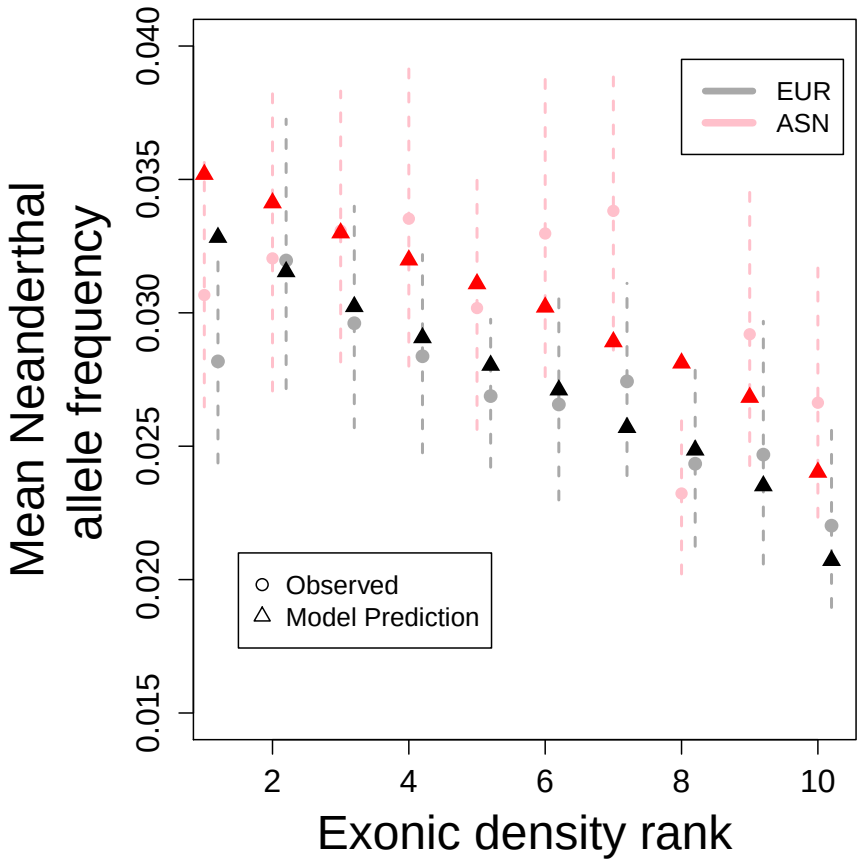

Figure S2.8: Fit between our estimates of p_t for bins of different exon density. Genomic regions with low exonic density (low exonic density rank) contain higher average Neanderthal allele frequency in both in Europeans (grey circle) and Asians (pink circle), a pattern recreated in our model. Dashed lines represent the 95% block bootstrap confidence intervals. The length of segments used to create the bins is 1.5 cM.

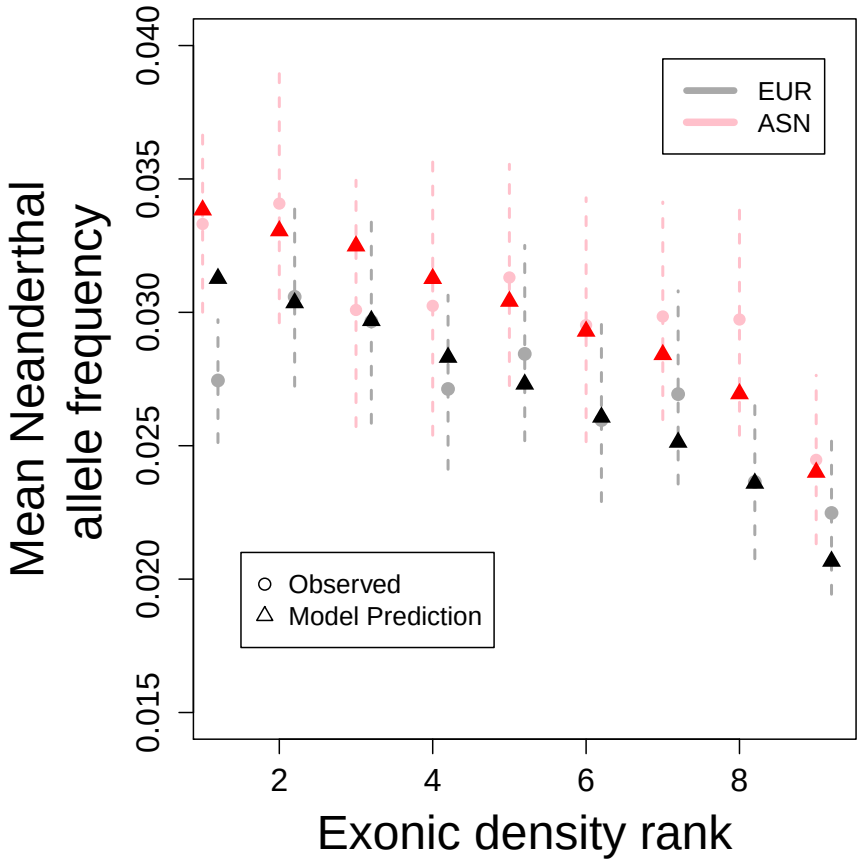

Figure S2.9: Fit between our estimates of p_t for bins of different exon density. Genomic regions with low exonic density (low exonic density rank) contain higher average Neanderthal allele frequency in both in Europeans (grey circle) and Asians (pink circle), a pattern recreated in our model. Dashed lines represent the 95% block bootstrap confidence intervals. The length of segments used to create the bins is 0.5 cM. There are 9 bins, rather than 10 bins, in this figure because there are many 0.5cM bins with zero exonic sites. Therefore, we collapsed our results together into a smaller number of bins.