SUPPLEMENTARY INFORMATION - 3: Individual-based simulations

Ivan Juric, Simon Aeschbacher, Graham Coop

1 Introduction

Here we describe individual-based simulations to investigate whether the differences in Neanderthal and human population sizes can account for the selection coefficient (s) and the density of deleterious sites (μ) found in the main paper to be associated with Neanderthal introgression into anatomically modern humans.

2 Theoretical and Technical Implementation

2.1 Simulation details

We use forward simulations under a simple Wright-Fisher model of the split between AMH and Neanderthals. Generations are non-overlapping. We assume that population of size N_a and ancestral to Neanderthals and humans was at selection-drift equilibrium when split occurred. Starting from the split time, we simulate a Neanderthal population of size N_n and a human population of size N_h for a duration T_D generations until the time of Neanderthal-human admixture. We assume that the ancestral population had a similar effective size to the long term effective human population size and set $N_a = N_h$.

We consider a single biallelic selected locus with two alleles A and a. We assume multiplicative fitness scheme, and set $w(A) = 1$ and $w(a) = 1 - s$. Each run is initiated with a different s drawn from a distribution of deleterious selective effects of new mutations f_s. We assume that initially the allele frequency was at the mutation-selection-drift equilibrium in the ancestral population. This allowed us to use the following fast procedure to initiate the frequency of deleterious allele in each simulation run.

For each run we first test whether the locus is either polymorphic or monomorphic in the ancestral population. The probability that a locus is polymorphic in ancestral population is equal to the probability that a is found at any frequency between $1/2N_a$ and $1 - 1/(2N_a)$, which in the diffusion limit is:

$$P(\text{site is polymorphic}) = 4N_a u \int_{1/2N_a}^{1-1/2N_a} P(x, N_a, s, u)dx, \tag{1}$$

where $P(x, N_a, s)$ is the stationary density of the diffusion at frequency x. This density is proportional to equation 9.3.3 in Crow and Kimura [1970]:

$$P(x) \propto e^{-4N_a sx} x^{4Nu-1}(1-x)^{4Nu-1}, \tag{2}$$

where u is the per site per generation mutation rate from a to A and also from A to a. We assume that the value of u does not change after the split. If the locus is polymorphic in the ancestral population, we draw a frequency x for the polymorphic a allele from the stationary distribution (2), otherwise the initial allele frequency is set to zero. We then run a Wright-Fisher forward simulation T_D generations. During the course of simulation, we introduce a new allele at frequency $1/(2N)$ with probability $2N\mu$ per generation.

Each simulation run ends after T_D generations. In total we perform R independent runs. For each run we record the frequency of deleterious allele in Neanderthal and human populations at the end of the simulation. We also record d, the difference between the frequency of deleterious allele in Neanderthal

and human population, $d = freq(a_N) - freq(a_H)$. A $d = 1$ implies that deleterious allele is fixed in Neanderthals and lost in humans, while $d = -1$ implies the converse.

2.2 Simulation Parameters

The following parameters were kept constant: $N_h = N_a = 10000$, $T_D = 20000$. We simulated three different Neanderthal population sizes (N_n): 500, 1000 and 2000. Those values span a range of Neanderthal effective population sizes proposed and used by others [e.g. Pruefer et al., 2014, Vernot and Akey, 2015]. For $N_n = 500$ we also used three different mutation rates ,u, (1×10^{-8}, 2×10^{-8} and 3×10^{-8}), that span a range of plausible per-nucleotide mutations rates in humans [Segurel et al., 2014]. Note that our mutation rate should be thought of as rate of non-synonymous mutation in genic regions to make it an appropriate match for the parameters of Boyko et al. [2008]. For other N_n we kept $u = 2 \times 10^{-8}$, however using other values for u did not strongly change our conclusions. For each parameter set, we simulated 8 million runs. To select a selection coefficient for each simulation run, we use the distribution of scaled selection coefficients from Boyko et al. [2008]. We use their estimated gamma distribution for $2Ns$ (with parameters $\alpha = 0.184$ and $\beta = 8200$), where $N = 25636$ the effective population size in their model [Boyko et al., 2008].

2.3 Comparing empirical estimates to simulations results

To mimic our assumptions about the deleterious Neanderthal alleles introgressing into AMH, we consider all of the simulations runs in which a deleterious allele was fixed in Neanderthals but is absent in humans.

In the main text we estimate a selection coefficient (s in the main text) and the density of selected sites (μ). Both of these parameters are statements about the distribution of selection coefficients that distort allele frequencies at linked neutral Neanderthal alleles. Weakly selected alleles ($s \ll 1/(2N_h)$) will not have a strong effect on linked neutral alleles, and so will not affect our estimate of μ and s. We therefore compare the empirical estimates to our simulations truncating the simulations below a variety cutoffs for selection coefficients.

Let $g_{c,s}$ be the distribution of selection coefficients, greater or equal than some cutoff value c, obtained from the subset of simulation runs with fixed Neanderthal differences. For example, $g_{0.01,s}$ is the distribution of selection coefficients that are ≥ 0.01 in all runs that ended with deleterious allele fixed in Neanderthals and lost in humans. We denote by $\bar{s}_c$, the mean of $g_{c,s}$. We view $\bar{s}_c$ as the estimate of mean selection coefficient of deleterious Neanderthal alleles assuming that alleles with selection coefficient $\leq c$ cannot be detected by our method (i.e. that will not affect admixture levels in humans). For a given cutoff c we also define μ_c, which our approximation to the estimate of μ given that our method cannot detect alleles with selected coefficient less than c. We calculate μ_c as the fraction of runs in which the deleterious allele with selection coefficient $\geq c$ was fixed in Neanderthals and absent in humans.

3 Results

3.1 Frequency of deleterious allele at the end of simulation

We ran our simulations over a range of parameters, with largely similar results. We report a range of summaries in Table 3.1. For example for a $N_h = 10000$ (Human population) and $N_n = 2000$ (Neanderthal population) out of 8 million simulations 13623 times a deleterious allele was present at the end of simulation (time of admixture) in Neanderthal population and 13690 in Human population. When present, the average frequency of deleterious allele was 0.874 in Neanderthals and 0.910 in humans, with the deleterious alleles more often fixed in the Neanderthal population (12323) than in the human population (10457).

3.2 Estimate of selection coefficient and the density of selected sites

Figure S3.1 shows the distribution of the difference of the frequency of deleterious allele between Neanderthal and human populations (d) split by different bins of s for $N_n = 2000$. A value of $d = 0$ was

N_n	u	#apH	#apN	freqH	freqN	#fixH	#fixN	#newH	#newN
2000	2×10^{-8}	13623	13690	0.874	0.910	10457	12323	5	337
1000	2×10^{-8}	13646	13854	0.875	0.915	10501	12553	7	586
500	3×10^{-8}	19953	20007	0.875	0.949	15392	18866	12	1200
500	2×10^{-8}	13785	13779	0.869	0.940	10415	12840	8	806
500	1×10^{-8}	6984	6514	0.810	0.928	5301	5988	4	443

Table S3.1: Summary of individual-based simulations. N_n: Neanderthal population size; u: mutation rate; #apH/#apN: number of runs in which the deleterious allele was present in Human/Neanderthal population at the end of the simulation; freqH/freqN: mean frequency of deleterious allele in runs where deleterious allele was present at the end of the simulation; #fixH/#fixN: number of times the deleterious allele fixed in the population; #newH/#newN: number of runs in which a fixed deleterious allele was initially not present in the population but appeared as a new mutation after the Human-Neanderthal split

the most common simulation outcome and we omitted this uninteresting case since it was several orders of magnitude larger than all other cases. A $d = 0$ for the most part indicates the loss or fixation of deleterious allele in both populations, but also (very rarely) the same frequency of deleterious allele at the end of a run in both populations.

When s is small (bottom histogram), $\ll 1/N_h$, Figure S3.1 shows that the deleterious allele is slightly more likely to be fixed in Neanderthals and lost in humans ($d = 1$) and fixed in humans and lost in Neanderthals ($d = -1$). This happens because selection is slightly more efficient at removing deleterious alleles from larger human population. We observe larger effects of the increased efficiency of selection to remove deleterious allele for larger selection coefficients (middle and top histograms). As selection becomes stronger the discrepancy between $d = 1$ and $d = -1$ becomes larger because such deleterious alleles are more rapidly removed from human population. Figure S3.1 also suggests that $1 \times 10^{-5} < s_{sim} < 1 \times 10^{-3}$ is the range of selection coefficients for which we are likely to observe substantially more deleterious alleles fixed in Neanderthals and lost in humans in our simulations. Our genome-wide estimate of mean selection coefficient for EUR and ASN populations is within this range. As expected this corresponds to selection coefficients in the range of $1/N_h < s < 1/N_n$, where the efficacy of selection against weakly deleterious in Neanderthals reduced compared to that in humans.

Figure S3.2 shows on the x axis the value of $\bar{s}_{c,g}$ for values of c ranging from 0 to 1×10^{-3} by 1×10^{-4} versus μ_c (see "Comparing empirical estimates to simulations results" section for details). Our estimates agree with the order of magnitudes in the simulations and overlap when Neanderthal population size was small ($N_n = 500$). Note also that our estimates of μ and s may represent over-estimates, as we include only exonic sites and so nearby functional non-coding sites may inflate our estimates of the density and strength of selection (as discussed in the main text). The reasonable agreement between these simulations and parameter estimates from our model suggests that it is quite plausible that nearly neutral alleles make up the bulk of deleterious introgressed Neanderthal alleles.

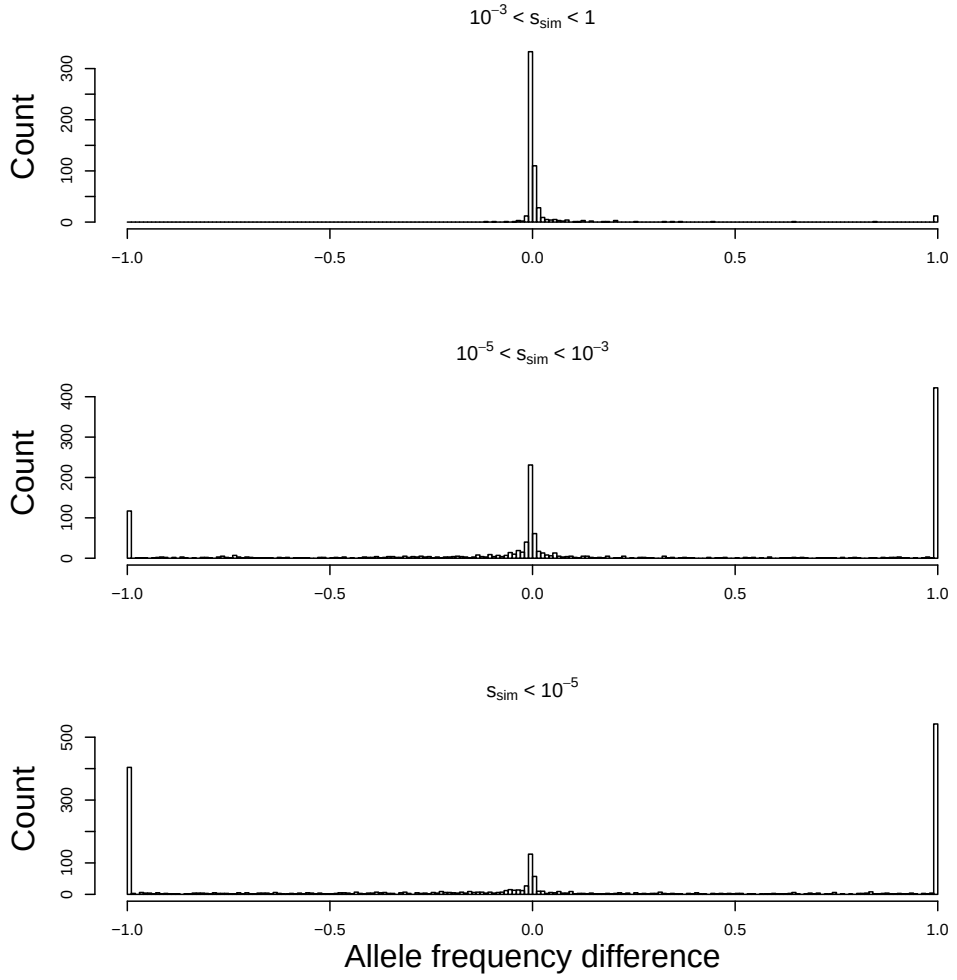

Figure S3.1: Differences in the frequency of deleterious allele between Neanderthal and human population split by the range of selection coefficients. Positive values indicate higher frequency of deleterious allele in Neanderthal population and negative values mean that deleterious allele was at higher frequency in human population than in Neanderthal. Other parameters: $N_n = 500$, $u = 2 \times 10^{-8}$

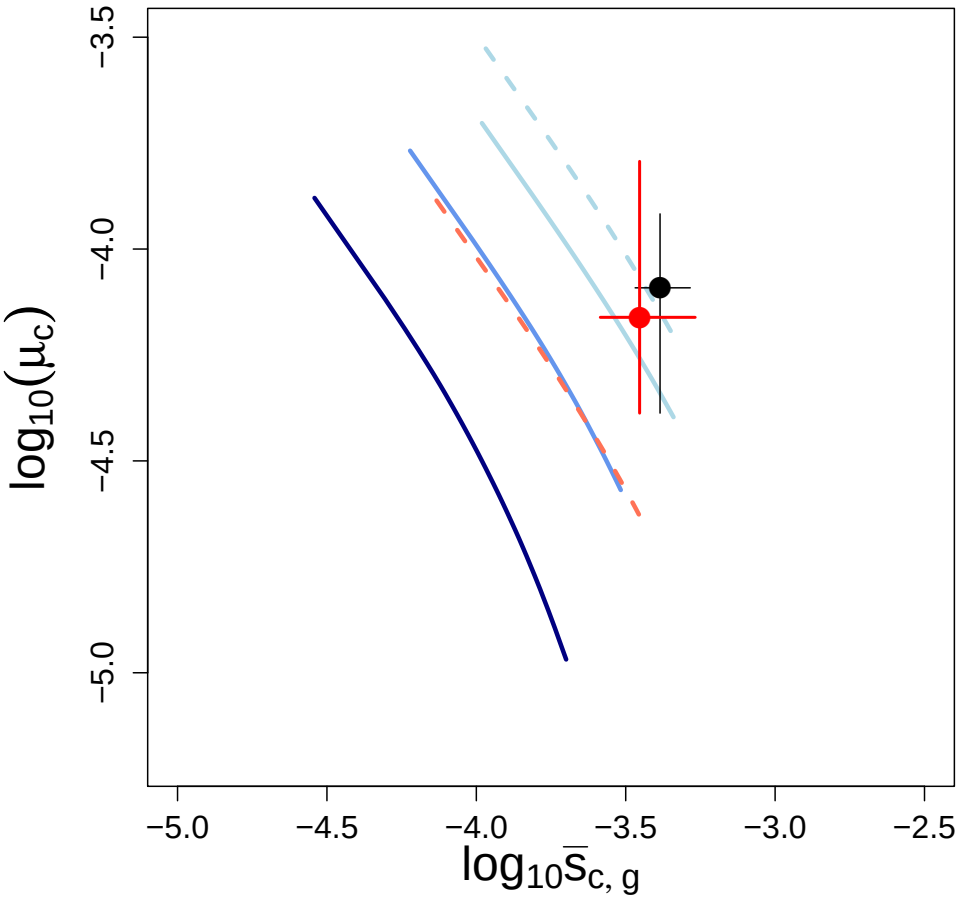

Figure S3.2: Mean selection coefficient ($\bar{s}_{c,g}$) versus μ_c across a range of cutoffs. Blue full lines represent the simulation results for our three different Neanderthal sizes, with the darkest line corresponding to $N_n = 2000$ and the lightest $N_n = 500$. The light blue dashed line represents simulation set in which $N_n = 500$ and $u = 3 \times 10^{-8}$ while the light red dashed line corresponds to $N_n = 500$ and $u = 1 \times 10^{-8}$. Lastly, the red points represent the estimates for EUR and ASN populations and red lines the appropriate 95% confidence intervals.

References

James F Crow and Motoo Kimura. An introduction to population genetics theory. *An introduction to population genetics theory.*, 1970.

Kay Pruefer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Helene Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Paeaebo. The complete genome sequence of a Neanderthal from the Altai Mountains. *NATURE*, 505(7481):43+, JAN 2 2014. ISSN 0028-0836. doi: {10.1038/nature12886}.

B. Vernot and J. M. Akey. Complex history of admixture between modern humans and Neandertals. *Am. J. Hum. Genet.*, 96(3):448–453, Mar 2015.

Laure Segurel, Minyoung J. Wyman, and Molly Przeworski. Determinants of Mutation Rate Variation in the Human Germline. In *ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS, VOL 15*, volume 15 of *Annual Review of Genomics and Human Genetics*, pages 47–70. ANNUAL REVIEWS, 2014.

Adam R. Boyko, Scott H. Williamson, Amit R. Indap, Jeremiah D. Degenhardt, Ryan D. Hernandez, Kirk E. Lohmueller, Mark D. Adams, Steffen Schmidt, John J. Sninsky, Shamil R. Sunyaev, Thomas J. White, Rasmus Nielsen, Andrew G. Clark, and Carlos D. Bustamante. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLOS GENETICS*, 4(5), MAY 2008. ISSN 1553-7390. doi: {10.1371/journal.pgen.1000083}.