Supporting Information

Eriksson and Manica 10.1073/pnas.1200567109

SI Text

Analyses of Candidate Regions for Gene Flow from Neanderthals. The original publication of the draft Neanderthal genome (1) included two different metrics to describe the degree of shared polymorphism between Neanderthals and different modern human populations. We investigated the effect of ancient population structure on both metrics. The effect on the *D* statistics [supplementary online material (SOM) 18 in ref. 1] is described in the main text. Here we describe the effect on the metric presented in SOM 17 of ref. 1, which focuses on ancestral and derived SNPs in sections of the human genome that have large times to most recent common ancestor (TMRCA).

SI Materials and Methods

We have attempted to replicate the analysis of Green et al. (1) as closely as possible within our model (see Fig. S5 for an outline of the main steps). The analysis compares the Neanderthal genome to African (AFR) and out-of-Africa (OOA) samples (individuals with European or Asian ancestry). Candidate regions were found by looking for regions with high TMRCA in the OOA sample compared with the TMRCA of the AFR sample. TMRCA was calculated from sequence data in 50-kb bins, sliding across the genome in steps of 10 kb, by first constructing an unweighted pair group method with arithmetic mean (UPGMA) tree (2) for the bin and then taking the average of the number of mutations from each tip of the tree to its root (the tips of the tree correspond to the observed sequences). Let S_T denote the ratio of TMRCA for the OOA sample to the TMRCA for the AFR sample for a given 50-kb bin. A region was then classified as "candidate" if six consecutive bins all had S_T values in the top 0.5% of all S_T values in the genome. Hence, candidate regions are 100 kb long.

Within each candidate region, tag SNPs were identified as follows. First, a UPGMA tree was constructed for the joint AFR and OOA sample. In this tree, tag SNPs are mutations on the root lineages, identified using parsimony, such that they separate a clade containing only lineages from the OOA sample from a joint OOA/AFR clade (Fig. S4, step 2). Finally, the tag SNPs were classified with respect to whether the OOA allele matches the Neanderthal (M for match or N for nonmatch), and whether the Neanderthal allele is derived (D) or ancestral (A), giving four categories: AM, AN, DM, and DN (Fig. S2, step 3).

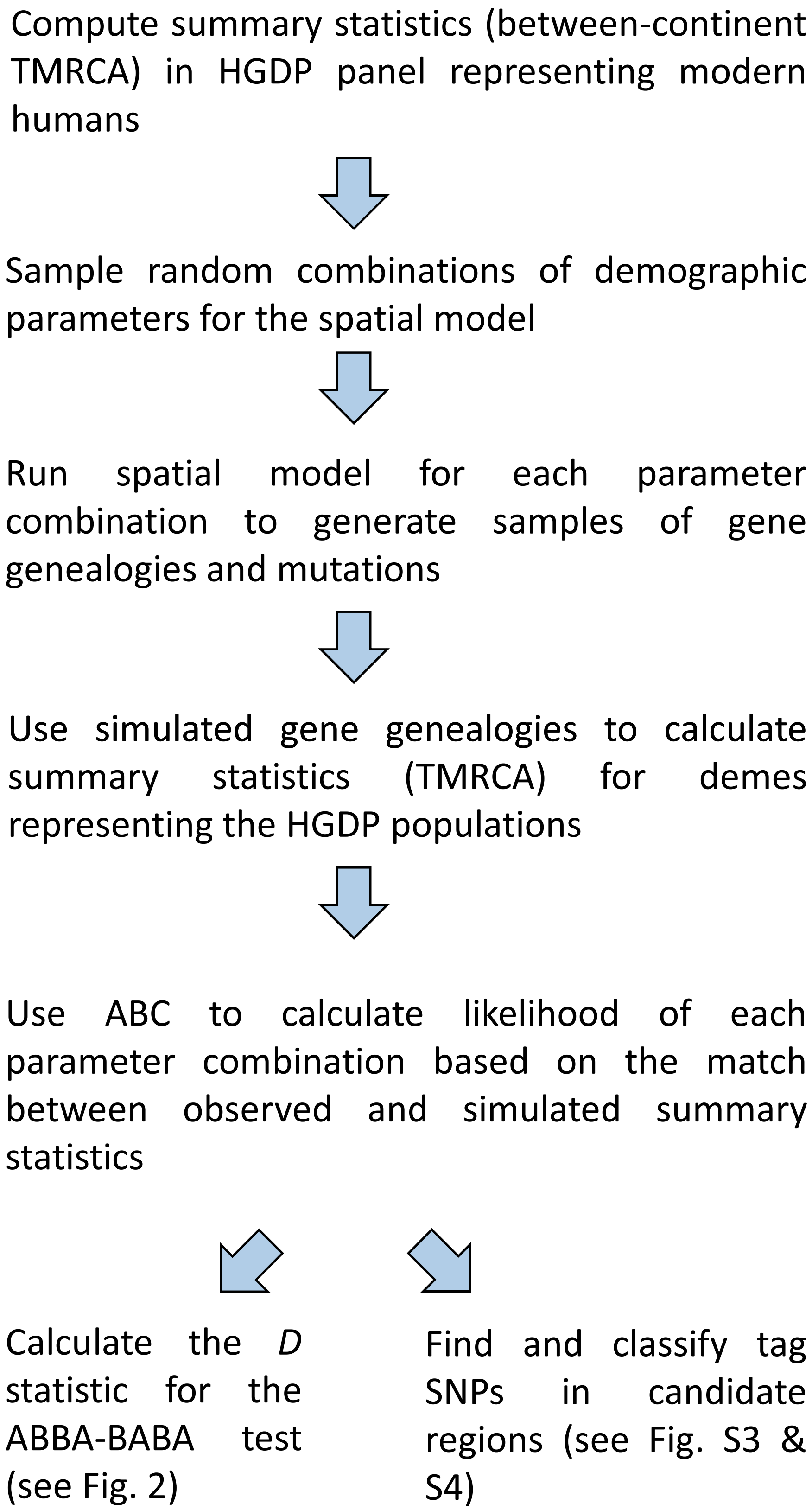
For each of the parameter combinations selected by ABC (*Materials and Methods*), we generated 150,000 100-kb regions, corresponding to approximately five full genomes for each individual and parameter combination. We attempted to match the sample design of Green et al. (1) as closely as possible within our model. When generating sample sequences from our models, we took 24 individuals from randomly chosen demes in the range 70–120 (corresponding to Europe), 24 individuals from demes 130–170 (East Asia), and 23 samples from Africa. Because the American African sample used by Green et al. (1) was admixed with 80% African and 20% European ancestry, for each 46 gamete in the African sample we picked it with probability 0.8 from the range 1–10 (sub-Saharan Africa) and with probability 0.2 from a random deme in the European range. For the ABBA-BABA analysis, we placed the Neanderthal on the northern branch of the stepping-stone model (which became separated from the African at the split between humans and Neanderthals) in the deme corresponding to the distance between their origin in sub-Saharan Africa and the Vindija cave in Croatia, where the Neanderthal sample used in Green et al.'s (1) analysis was found. We simulated unlinked 100-kb regions by generating the gene genealogy of the sample as described in the previous section. We then generated mutations in the gene genealogy according to the Jukes–Cantor model (assuming a split with chimpanzees 6 Mya) with mutation rate 2.5×10^{-8} per gamete per generation (3). Using the procedure described above, we identified candidate regions and their tag SNPs from the simulated samples. Finally, we generated the distribution of fraction of matching for derived (DM vs. DN) and ancestral (AM vs. AN) as follows. First, for each of the selected demographic scenarios we used bootstrapping to generate the corresponding distributions of fraction of matching tag SNPs (for ancestral and derived separately) in a sample of 166 randomly chosen tag SNPs. Second, we averaged the distributions over the demographic scenarios weighted by likelihood as estimated by ABC (*Materials and Methods*).

SI Results

In the real data, there is an excess of matching vs. nonmatching positions (90.5% of derived SNPs are matching, and 69.5% of ancestral SNPs are matching, as obtained from table 5 in ref. 1). Green et al. (1) used a model with no population structure within continents to represent the split of Neanderthals from ancestors of anatomically modern humans and the latter's subsequent spread out of Africa. With no admixture between Neanderthals and modern humans, this model failed to capture the excess of matching to nonmatching (15.6% of derived and 50.7% of ancestral predicted as matching). However, introducing 1–4% of admixture just after modern humans' spread out of Africa raised the fraction of matching to 77.5% among derived SNPs (but gave only 38.3% matching of ancestral SNPs).

When we considered the best 0.05% of our parameter combinations, weighted by the likelihoods estimated by ABC, our predictions were compatible with the observed patterns. Indeed, the most likely proportions of matching for both derived and ancestral SNPs were closer to the observed values than the predictions of the models with and without admixture presented in Green et al. (1), with 83.4% of derived and 67.1% of ancestral SNPs matching (Fig. S4 *A* and *B*, respectively).

1. Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328:710–722.
2. Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 28:1409–1438.
3. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.



Compute summary statistics (between-continent TMRCA) in HGDP panel representing modern humans

Sample random combinations of demographic parameters for the spatial model

Run spatial model for each parameter combination to generate samples of gene genealogies and mutations

Use simulated gene genealogies to calculate summary statistics (TMRCA) for demes representing the HGDP populations

Use ABC to calculate likelihood of each parameter combination based on the match between observed and simulated summary statistics

Calculate the *D* statistic for the ABBA-BABA test (see Fig. 2)

Find and classify tag SNPs in candidate regions (see Fig. S3 & S4)

Fig. S1. Schematic representation of the logic behind model fitting.

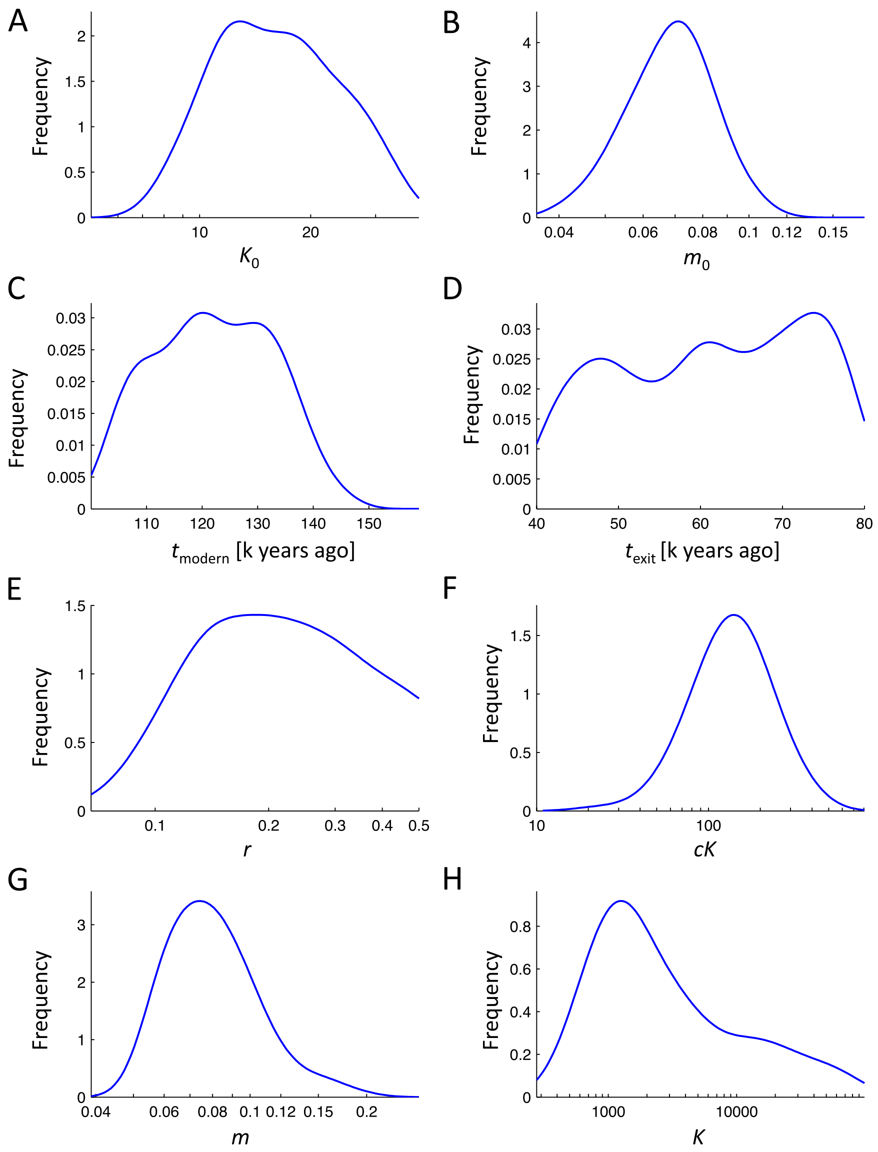


Fig. S2. Posterior distributions of the demographic parameters based on ABC. (*A*) Ancient carrying capacity, K_0; (*B*) migration rate, m_0; (*C*) timing for the transition from ancient to modern demography, t_{modern}, in kya; (*D*) timing for the expansion of modern humans out of Africa, t_{exit}, in kya; (*E*) modern growth rate, r, during the out-of-Africa expansion; (*F*) modern effective founder population size, cK; (*G*) modern migration rate, m; and (*H*) modern carrying capacity, K.

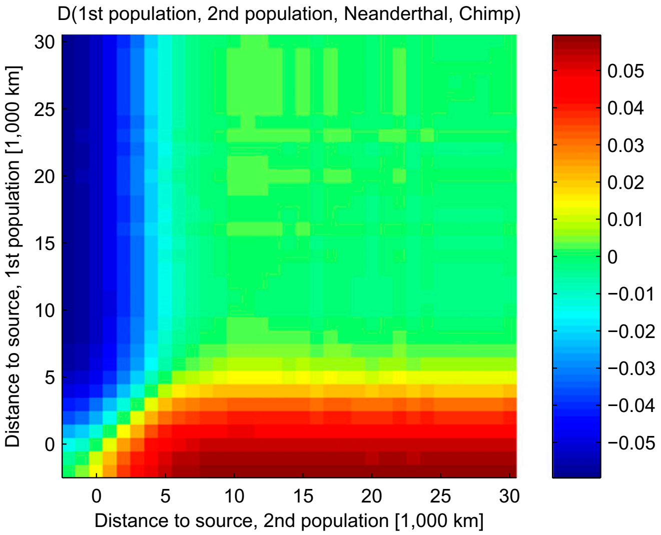


Fig. S3. *D* values for every pair of populations, averaged over the best-fitting scenarios.

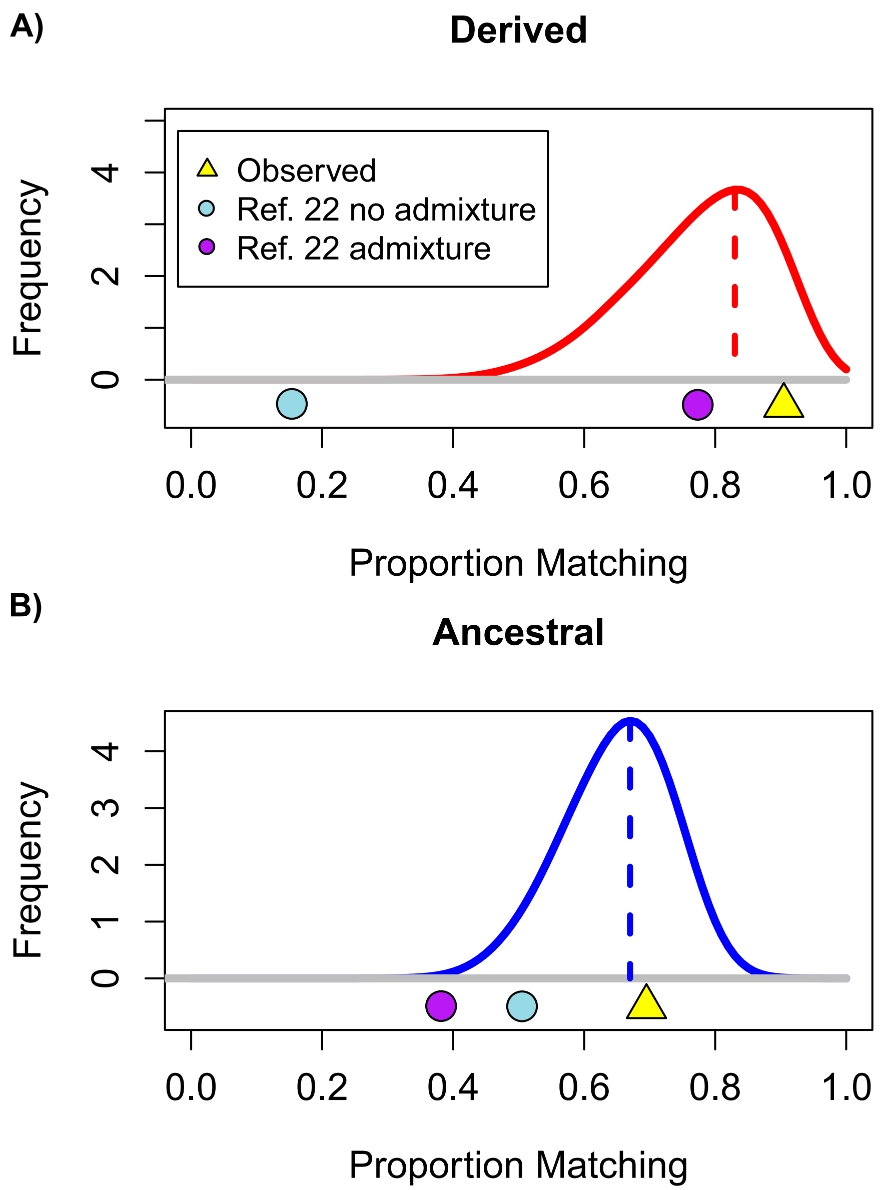
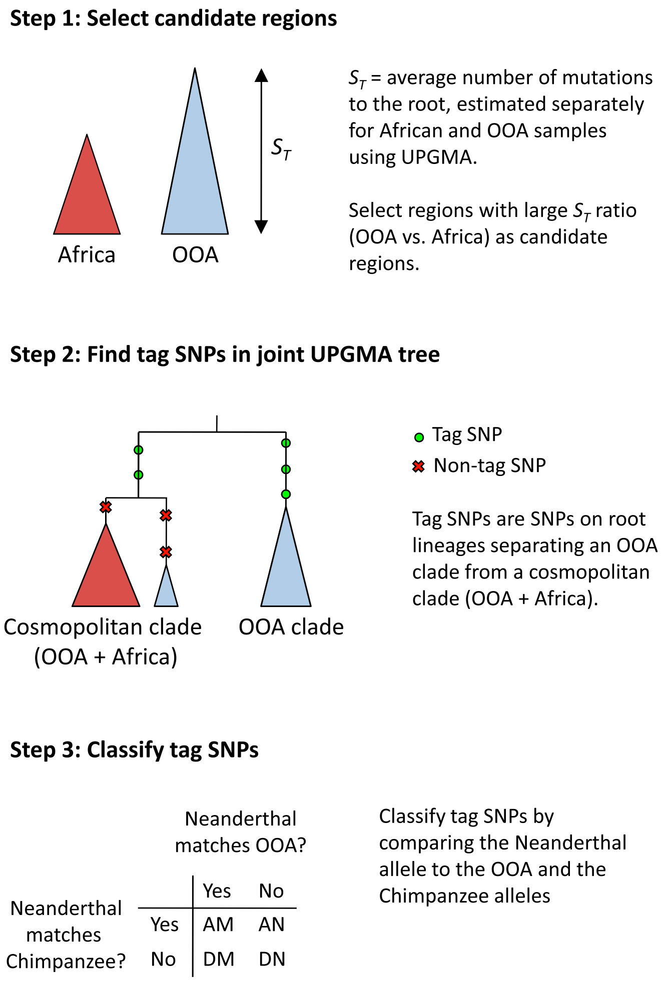


Fig. S4. The proportion of Neanderthal alleles matching those of Eurasians, separated into derived (*A*) and ancestral (*B*). In each panel we show the actual data (yellow triangle) and the proportions predicted by the Green et al. (1) model without (light blue circle) and with admixture between Neanderthal and Eurasians (purple circle). Results from our spatially structured model are presented as frequency plots of predicted values (red line for derived, and blue for ancestral), with a vertical dashed line representing the most likely fit.



Step 1: Select candidate regions

S_T = average number of mutations to the root, estimated separately for African and OOA samples using UPGMA.

Select regions with large S_T ratio (OOA vs. Africa) as candidate regions.

Step 2: Find tag SNPs in joint UPGMA tree

● Tag SNP

✖ Non-tag SNP

Tag SNPs are SNPs on root lineages separating an OOA clade from a cosmopolitan clade (OOA + Africa).

Step 3: Classify tag SNPs

		Neanderthal matches OOA?	
		Yes	No
Neanderthal matches Chimpanzee?	Yes	AM	AN
	No	DM	DN

Classify tag SNPs by comparing the Neanderthal allele to the OOA and the Chimpanzee alleles

Fig. S5. Schematic representation of the candidate region approach.