File S2 - Balancing selection under different demographic scenarios

Method To investigate how balancing selection affects variation under demographic scenarios that approximate those of the populations we sampled, we simulated human evolutionary history using population sizes, growth rates, mutation rates, and recombination rates previously estimated for human populations (Gravel *et al.* 2011).

We used the *msms* program (Ewing and Hermisson 2010) to simulate evolution under both neutrality and selection. Our simulations of selection assumed two models of heterozygote advantage (subsequently referred to as overdominance): shared overdominance and divergent overdominance, described below. For each simulation, an ancestral population evolved under overdominance for 2,000 generations. The population was then split into two daughter populations which were allowed to evolve without gene flow for t generations, after which the simulation was stopped and F_{ST} was estimated between the population pair. We ran independent simulations for values of t in the range between 50 to 2,500 generations, which encompasses population divergence times compatible with the Out of Africa model of human evolution. We assumed that both populations experienced exponential growth after splitting from their common ancestor, which had effective size of one thousand individuals ($N_0 = 1000$). Growth rates (g) were defined so as to satisfy different values of fold-change in population size, following the equation: $N_t = N_0 \exp(\frac{gt}{4N_0})$, where t is the time elapsed since beginning of growth in generations and N_0 is the initial effective population size. Growth was assumed to start at the exact moment that the populations split from one another. Sequences were 5,000 bases long, and recombination occurred at a rate of 10^{-8}. We chose a value of N_0 of the same order as the effective population size of non-African populations. The fold-change values (and corresponding growth rates) were kept below an 8-fold change, since we documented that this change is sufficient to substantially change the impact of selection on F_{st} in the case of constant overdominance (Figure S2). The length of the genes was based on the median size of classical HLA loci (Lenz *et al.* 2016) and the selection coefficients were chosen to span a range approximating the highest found for HLA (Yasukochi and Satta 2013).

We used this simulation framework to explore three scenarios: (a) neutral evolution; (b) a "shared overdominance scenario", where the two daughter populations experience identical selection; (c) a "divergent overdominance scenario", where the fitness of the homozygotes differs between the daughter populations.

To simulate overdominance, we place a biallelic variant under heterozygote advantage in the first position of the sequence. The selective regime is defined by both the selection coefficient (s) as well as the resulting equilibrium frequency attained (denoted by f_{eq}, representing the equilibrium frequency of the most common allele). We explore several combinations of f_{eq} (0.5, 0.7 and 0.9) and s (0.01, 0.05, 0.1). Note that the same f_{eq} can be reached with different values of s, but the trajectory to equilibrium will be faster when selection is stronger. All sites except the first one evolved under neutrality.

Specifically, the overdominant fitness values used in the simulation were written using the Malthusian formulation as required in the *msms* software (Ewing and Hermisson 2010), where $\omega_{aa} = 2N_e$, $\omega_{Aa} = 2N_e(1 + s)$ and $\omega_{AA} = 2N_e(1 + hs)$ with $h = 2 - 1/f_{eq}$ assuming a to be the least frequent allele and $f_{eq} \geq 0.5$ the frequency of the allele A.

For the "divergent ovedominance" scenario, after the split one population continues to experience overdominance under the same parameters as the ancestral population, and the second population experiences a new regime, with equilibrium allele frequencies differing from that of the previous regime by a value of δ. For example, if overdominance was chosen such that the expected equilibrium frequency in population 1 was p=0.5 and q=0.5, with a δ of 0.2, population 2 was placed in a novel selective regime where the equilibrium frequencies were p=0.3 and q=0.7.

Results Each point in Figure S2 is the average F_{ST} for all segregating sites, taken over 2000 simulations under shared overdominance ($\delta = 0$). F_{ST} values were estimated as the ratio of average variance components over all the SNPs in the simulated sequence. As expected, we observe a gradual increase in F_{ST} for both selected and neutral markers with increasing split times. Notice that for all growth rates differentiation increases at a lower rate when there is overdominance. Higher population growth results in lower differentiation, under neutral and selective scenarios.

Next, we simulated divergent overdominance, with each pair of populations splitting from a source population with $f_{eq} = 0.5, 0.7$ or 0.9 and suffering a shift in equilibrium allele frequencies of magnitude $\delta = 0, 0.2$ or 0.4. Figure S3 shows the difference in $\mathrm{F_{ST}}$ under neutrality and overdominance (ΔF_{ST}) for three shared overdominance (first row) and six divergent overdominance (second and third rows) scenarios. Only under divergent overdominance with $\delta \geq 0.2$ and $s > 0.05$ is $\mathrm{F_{ST}}$ higher at HLA than in neutral sites ($\Delta F_{ST} < 0$) for recent split times, similar to the data presented in the main text (Figure 6). The increase in $\mathrm{F_{ST}}$ for recent split times is attained for a wide range of f_{eq} of the source population.

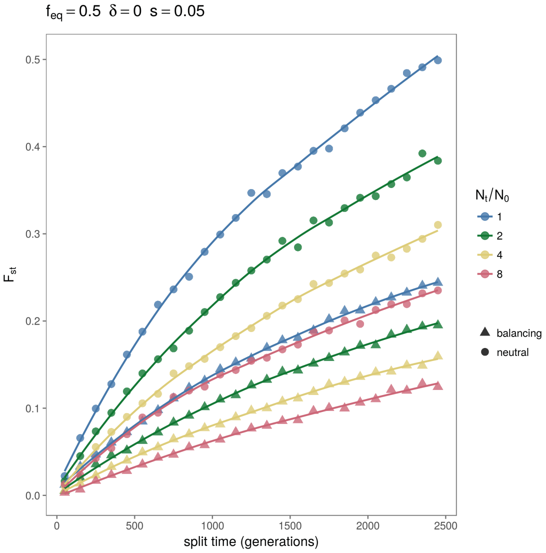

Figure S2 F_{ST} as a function of population split time under neutrality and shared overdominance. Population size expansion rates (N_t/N_0, varying from 1 to 8-fold change) are indicated in different colors. Simulated equilibrium frequency (f_{eq}) was 0.5 and selective coefficient (s) was 0.05.

Conclusion None of the combinations of population size change and split times simulated under shared overdominance could reproduce the observation that in pairs of populations with low divergence, the HLA genes show higher population differentiation than putatively neutral regions. Only when simulations included a shift in the overdominant selection equilibrium frequency (divergent overdominance), was this observation reproduced. Our simulation results show that a regime of divergent overdominance can reproduce the observation of higher population differentiation at the HLA genes than in neutral regions at low divergence times, and lower population differentiation at the HLA genes than in neutral regions at high divergence times (see also the Discussion section in the main text).

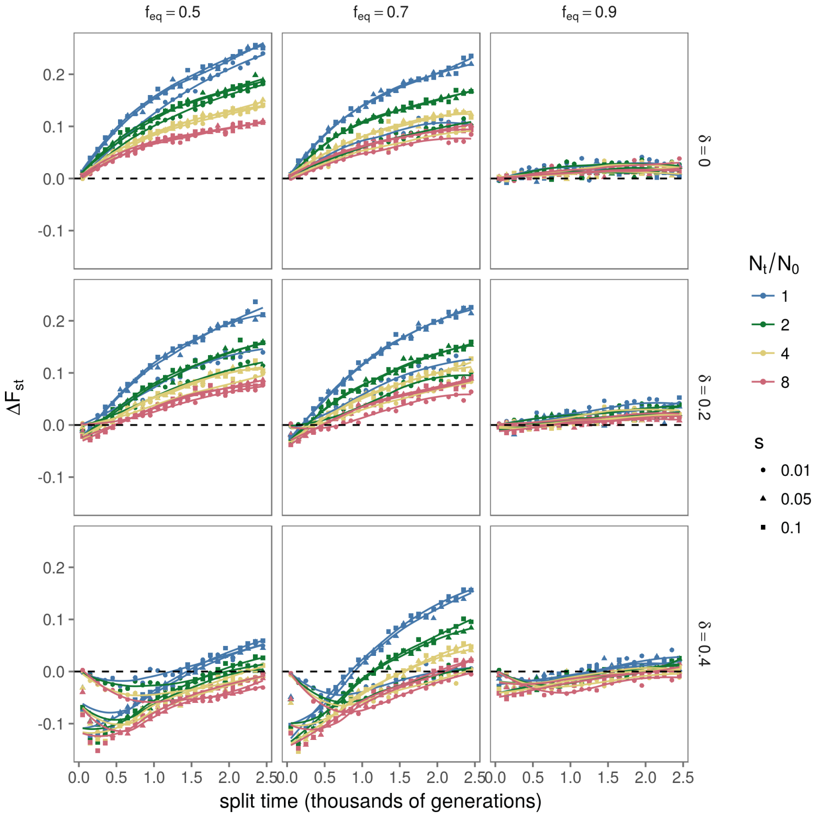

Figure S3 Difference in F_{ST} under selection and neutrality, for different selective scenarios and allele frequencies prior to population splits. ΔF_{ST} is the difference between F_{ST} under neutrality and F_{ST} under overdominance for simulated scenarios. Simulations varied population split times, population size fold change (N_t / N_0), selective coefficient (s), equilibrium frequencies (f_{eq}), and difference of equilibrium frequencies after split (δ).

Literature Cited

Ewing, G. and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064.

Gravel, S., B. M. Henn, R. N. Gutenkunst, a. R. Indap, G. T. Marth, *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences .

Lenz, T. L., V. Spirin, D. M. Jordan, and S. R. Sunyaev, 2016 Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. Molecular Biology and Evolution **33**: 2555–2564.

Yasukochi, Y. and Y. Satta, 2013 Current perspectives on the intensity of natural selection of MHC loci.