Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection

Tobias L. Lenz[1,2], Victor Spirin[1], Daniel M. Jordan[1] & Shamil R. Sunyaev[1,3]

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[2] Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany
[3] Program in Medical and Population Genetics, The Broad Institute, Cambridge, MA 02142, USA

Content: **Page**

Supplementary Notes

Supplementary Note S1: Distribution of derived allele frequencies within subpopulations

The ESP6500 release contains data from two human subpopulation, African Americans (AAs) and European Americans (EAs). The trend for elevated allele frequencies of deleterious variants in genes of the MHC regions holds also within subpopulation of the ESP data set. For African Americans (N = 2,203), the shift in the site frequency spectrum (SFS) of probably damaging variants is statistically significant (Wilcoxon rank-sum test, $P < 0.001$; **Fig. S4a**). In loss-of-function (LOF) variants, the frequency shift is not significant (Wilcoxon rank-sum test, $P = 0.52$), but follows the same trend (**Fig. S4b**): Across a range of frequency thresholds, the fraction of LOF alleles with that or higher frequency is consistently larger in the MHC region (**Tab. S1**). For European Americans (N = 4,300), the SFS shift in both probably damaging and LOF variants is statistically significant (Wilcoxon rank-sum test, both $P < 0.01$; **Fig. S4c/d**).

Supplementary Tables

Supplementary Table S1: Distribution of loss-of-function variants in African Americans

Fraction and number of loss-of-function (LOF) variants above different lower frequency thresholds in the African American subpopulation of the ESP6500 data.

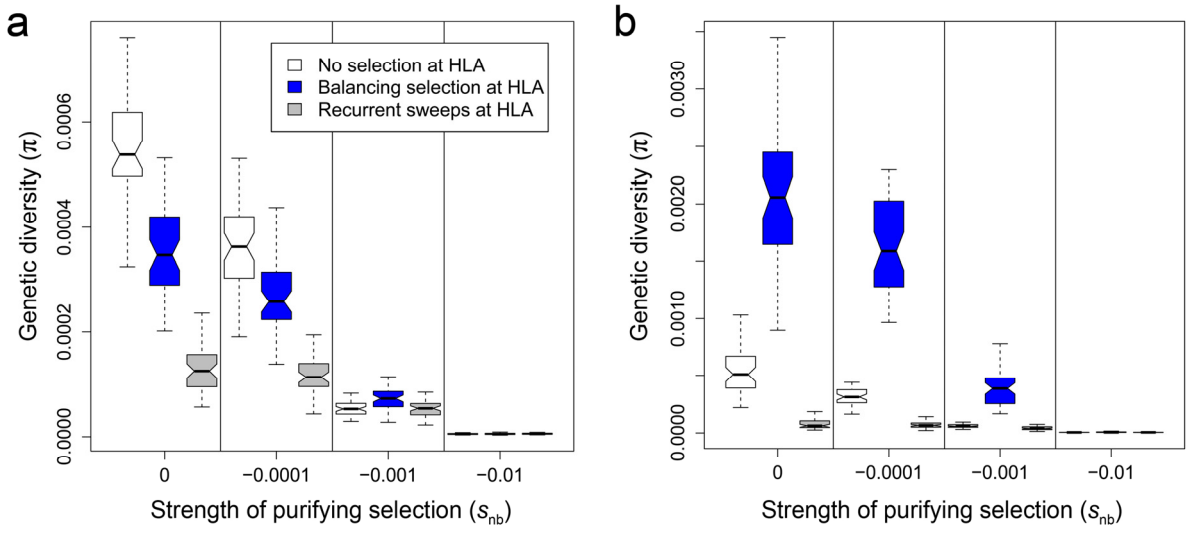
Lower frequency thresholds	**Complete exome**		**MHC region only (without classical HLA loci)**	
	Fraction [%]	#LOF variants	Fraction [%]	#LOF variants
0.01	1.9	296	2.8	2
0.005	2.7	418	2.8	2
0.001	6.6	1018	8.3	6
0.0005	10.0	1546	18.1	13

Supplementary Table S2: Genes in the MHC region carrying potentially deleterious variants with elevated frequencies

Listed are the 25 genes in the MHC region with the highest average derived allele frequency (DAF) of potentially deleterious variants in the ESP6500 data, including the number and frequency of probably damaging (based on PolyPhen-2 prediction) and loss-of-function variants. Also listed for each gene are relevant diseases reported in NHGRI GWAS Catalog and NIH Genetic Association Database. Note that the information from these databases does not always represent independent and/or causal effects. For some genes only a subset of reported diseases is listed. Gene position on chromosome 6 is based on GRCh37. Genes are sorted according to average DAF of probably damaging variants.

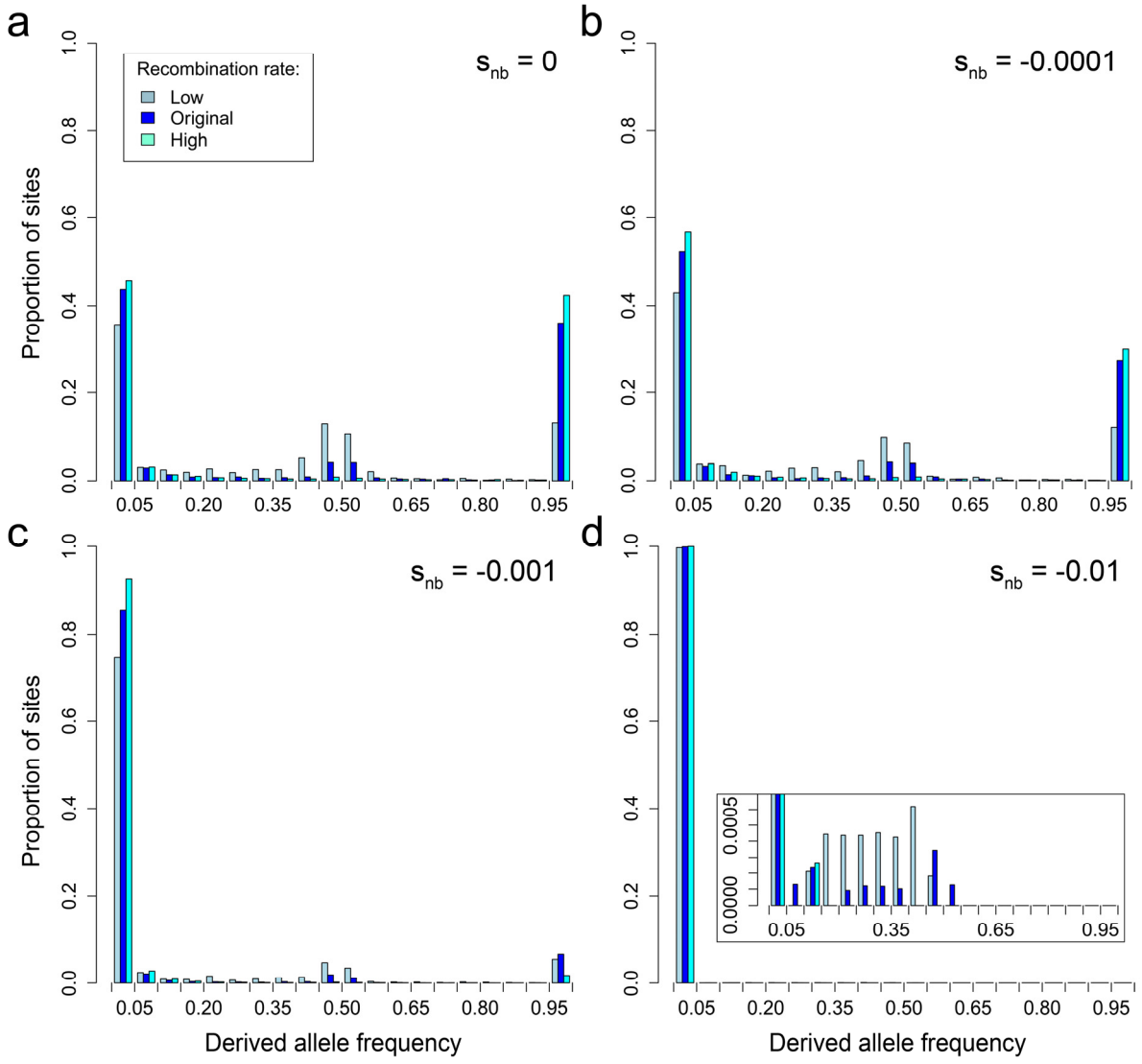
Gene	Position (first SNP)	Probably damaging variants		Loss-of-function variants		Diseases (NIH Genetic Association Database)	Diseases (NHGRI GWAS catalog)
		# SNPs	av. DAF [%]	# SNPs	av. DAF [%]		
PPT2	chr6:32121361	5	17.973	0	0.000	Pulmonary function, type 1 diabetes	Prostate cancer, pulmonary function
MICA	chr6:31371397	6	9.715	0	0.000	Behcet's disease, type 1 diabetes	Hepatocellular carcinoma, HIV-1 control, rheumatoid arthritis, Behcet's disease
PSORS1C1	chr6:31097369	2	8.815	1	0.012	Multiple sclerosis, psoriasis	Toxic epidermal necrolysis, systemic sclerosis, Crohn's disease
MCCD1	chr6:31496751	3	5.415	0	0.000	AIDS progression, multiple sclerosis	AIDS progression
LY6G5B	chr6:31638728	6	2.878	0	0.000	unknown	none
GPANK1	chr6:31629995	7	2.805	0	0.000	unknown	none
DPCR1	chr6:30908806	14	2.158	1	0.008	Systemic lupus	HIV-1 control
CFB	chr6:31914024	9	1.992	2	0.010	Macular degeneration, systemic lupus	Age-related macular degeneration, prostate cancer
RNF39	chr6:30038875	3	1.503	1	0.008	AIDS progression, systemic lupus, multiple sclerosis	AIDS progression
PRRC2A	chr6:31590542	56	1.387	0	0.000	unknown	Schizophrenia
C6orf15	chr6:31079111	11	1.180	0	0.000	Leprosy, systemic lupus	Follicular lymphoma, Graves' disease
C6orf10	chr6:32260736	4	1.167	3	0.020	Multiple sclerosis, type I diabetes, diabetic nephropathies, Rheumatoid arthritis, Asthma	Asthma, multiple sclerosis, rheumatoid arthritis
CCHCR1	chr6:31110338	31	0.990	0	0.000	Psoriasis, ulcerative colitis, prostate cancer	Prostate cancer, multiple myeloma
VARS2	chr6:30882634	25	0.833	3	0.009	Leprosy, HIV-1 control	Ulcerative colitis, HIV-1 control
KIAA1949	chr6:30645012	11	0.787	2	0.020	unknown	none
TRIM15	chr6:30131468	6	0.775	1	0.122	Multiple sclerosis, Psoriasis	none
MDC1	chr6:30668226	39	0.767	1	0.008	unclear	Primary biliary cirrhosis
ZFP57	chr6:29640247	5	0.668	1	0.013	Nasopharyngeal carcinoma	none
NEU1	chr6:31827505	4	0.659	0	0.000	unknown	none
TRIM31	chr6:30071279	5	0.640	1	0.356	Cardiomegaly, psoriasis	None
TNXB	chr6:32010144	95	0.614	0	0.000	Macular degeneration, type 1 diabetes, systemic lupus, multiple sclerosis	Age-related macular degeneration, HIV-1 control, systemic lupus, prostate cancer
BTNL2	chr6:32362521	19	0.544	1	0.640	Leprosy, systemic lupus, multiple sclerosis, type 1 diabetes, ulcerative colitis, biliary cirrhosis	Asthma, lung adenocarcinoma, sarcoidosis, ulcerative colitis, multiple sclerosis, dementia, Crohn's disease, prostate cancer
C2	chr6:31868837	5	0.536	0	0.000	Macular degeneration, systemic lupus, C2-deficiency, hepatocellular carcinoma	Age-related macular degeneration, prostate cancer, schizophrenia, Crohn's disease, ulcerative colitis, systemic lupus
C6orf26	chr6:31730793	5	0.486	1	0.095		None
TAP1	chr6:32813421	22	0.485	4	0.014	Grave's disease, psoriasis, asthma, cervical cancer, cystic fibrosis, primary glaucoma	Breast cancer, nephropathy, Alzheimer's disease

Supplementary Figures



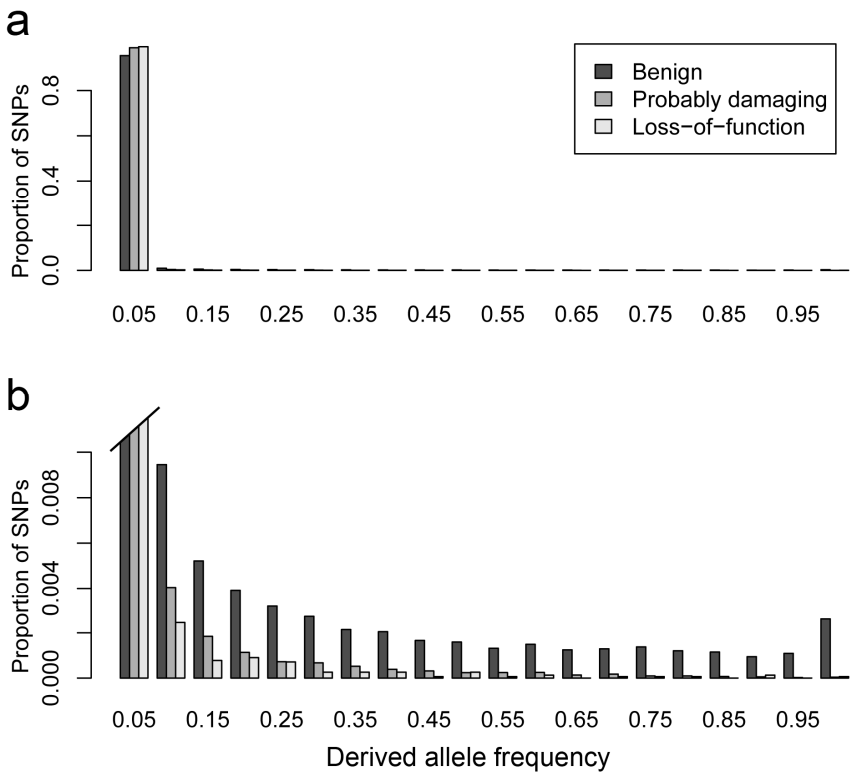
Supplementary Figure S1: Effect of different levels of recombination on simulated nucleotide diversity around the HLA

Nucleotide diversity along the regions surrounding the HLA gene are shown for simulations with (**a**) high (*r* = 4.4e-8) and (**b**) low (*r* = 4.4e-10) rates of recombination, in contrast to the original empirical level used in the main simulations (*r* = 4.4e-9). Same representation of HLA selection scenarios as in **figure 2c** in the main manuscript. Each simulation was replicated 50 times. Note the different y-axis scales.



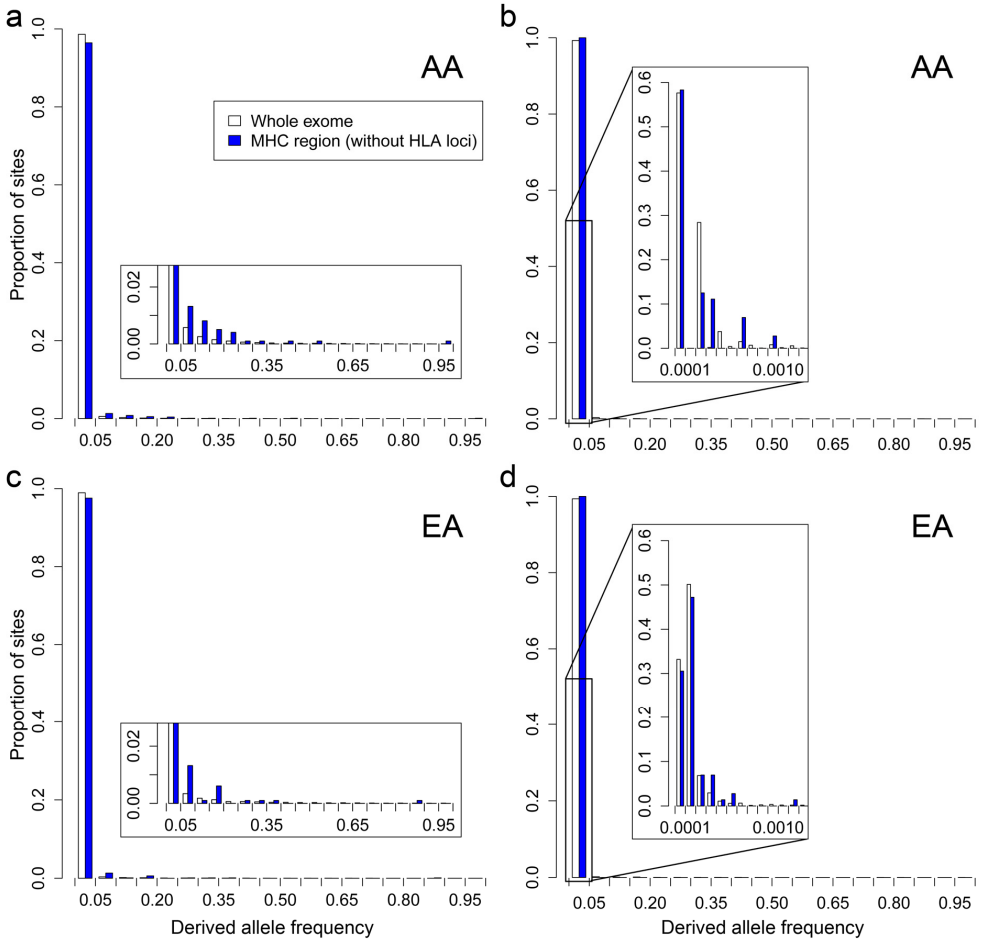
Supplementary Figure S2: Simulated site frequency spectrum of variants surrounding the HLA gene under balancing selection with different levels of recombination

Derived allele frequencies along the regions around the HLA gene are derived from simulations with three different levels of recombination (light blue: low recombination rate [r = 4.4e-10], dark blue: original [r = 4.4e-9], cyan: high [r = 4.4e-8]). Here the HLA gene evolved under balancing selection in all scenarios, while the variants in neighboring regions evolved (**a**) neutrally (s_{nb} = 0) or under co-dominant purifying selection with (**b**) s_{nb} = -0.0001, (**c**) s_{nb} = -0.001, or (**d**) s_{nb} = -0.01, respectively. Note the different y-axis scale in the zoomed inset of panel (**d**) for better visualization. Each simulation was replicated 50 times.



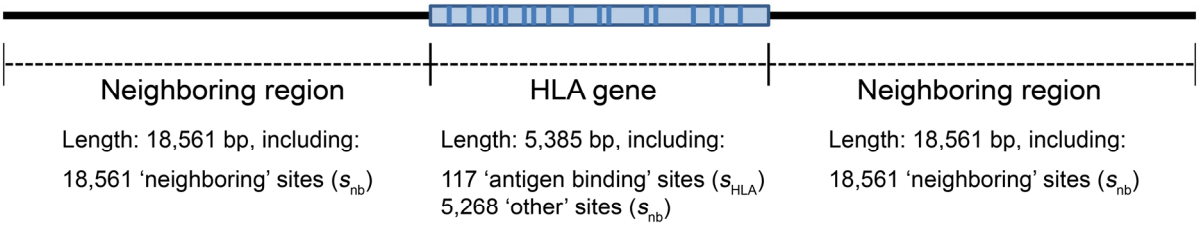
Supplementary Figure S3: Site frequency spectra of different missense variant categories across the entire exome

Distribution of exome-wide derived allele frequencies for missense variants predicted to be “benign” or “probably damaging” by PolyPhen-2, as well as for loss-of-function (stop-gain) variants. Allele frequency data from the ESP6500 exome sequencing dataset, shown (a) at full scale and (b) zoomed in to visualize rare variant classes.

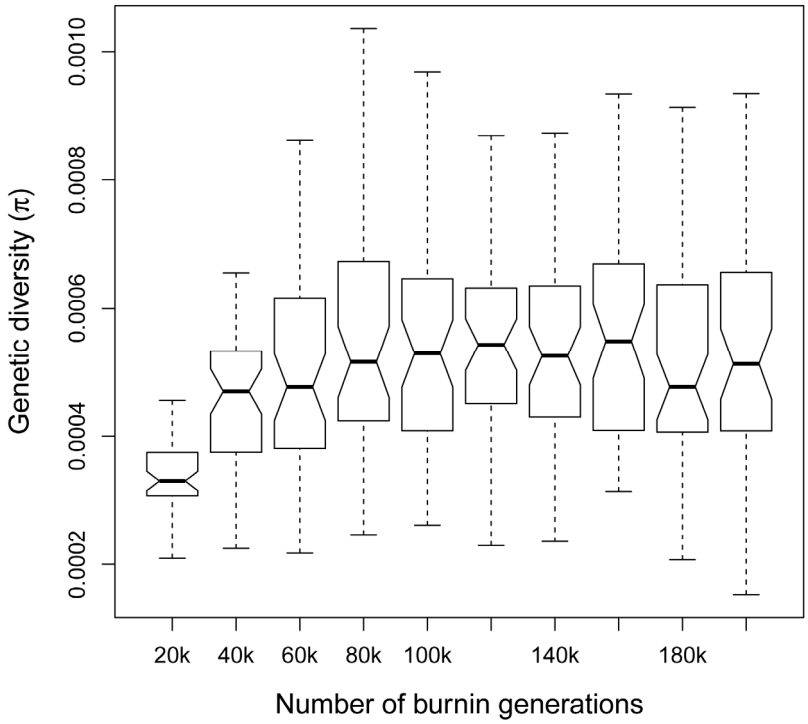


Supplementary Figure S4: Site frequency spectra (SFS) within subpopulations of the ESP6500 dataset

Distribution of derived allele frequencies within subpopulations are shown for potentially deleterious variants (‘probably damaging’ and loss-of-function) in African Americans (AA, a) and b) respectively) and in European Americans (EA; c) and d) respectively). Note the different x- and y-axis scales in the four insets.



Supplementary Figure S5: Schematic of the simulated HLA region
Shown is the model of the simulated HLA region as described in the methods. The total length of the simulated genome is 42,507 bp. The virtual HLA gene in the middle contains 117 sites that are evolving under the specified HLA selection scenario and are specified at the beginning of each simulation run (after burnin). It furthermore contains 5,268 sites that evolve under the same selection as the neighboring regions, but which are excluded from analyses, as we are only interested in genetic variation neighboring the HLA gene. The neighboring regions each contain 18,561 sites that evolve under different levels of purifying selection. For reasoning of the lengths of the different regions and a description of the different selection regimes see the original methods section.



Supplementary Figure S6: Simulated neutral genetic diversity after different burnin periods

Shown is the level of neutral genetic diversity achieved after different periods of burnin (20,000 – 200,000 generations, N = 10,000). The equilibrium level of 4Nμ was reached after approximately 80,000 – 100,000 generations. Boxplots show median, interquartile range, and extreme data points of 50 replicated simulations.