Cell Genomics, Volume *3*

Supplemental information

Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe

Gaspard Kerner, Anna-Lena Neehus, Quentin Philippot, Jonathan Bohlen, Darawan Rinchai, Nacim Kerrouche, Anne Puel, Shen-Ying Zhang, Stéphanie Boisson-Dupuis, Laurent Abel, Jean-Laurent Casanova, Etienne Patin, Guillaume Laval, and Lluis Quintana-Murci

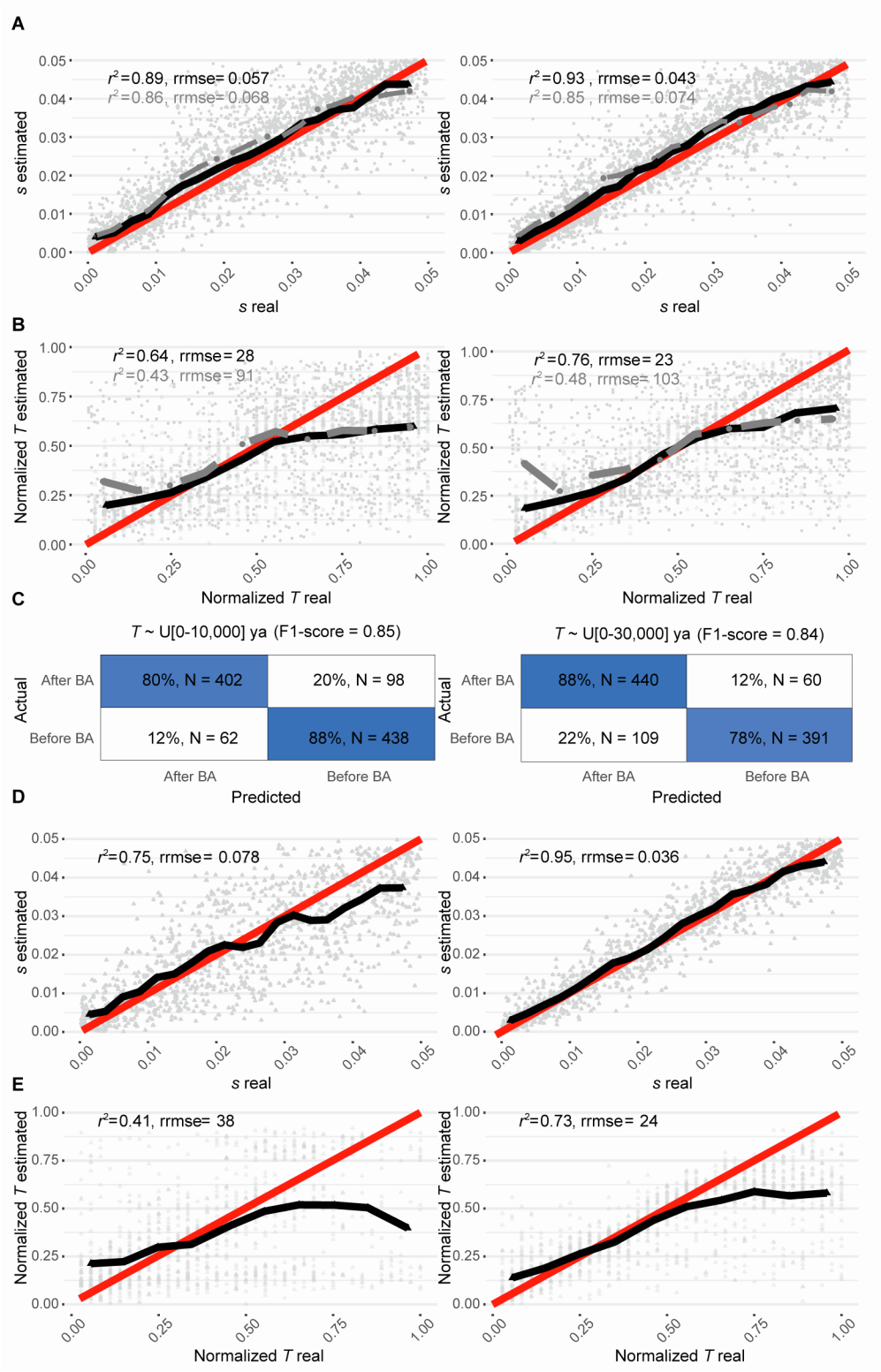

Figure S1. Cross-validation for the time of selection onset and selection intensity, Related to STAR Methods

Leave-one-out cross-validation for (A,D) the selection coefficient (*s*) or (B,E) the onset of selection (*T*), for either (A,B) positive or (D,E) negative selection. The left column shows analyses on variants for which the maximum of the lower bounds of the CI for the frequency of the variant across epochs was between 0 and 20% (**STAR Methods**) and the right column shows analyses on variants for which this maximum was between 20 and 60%. Each panel

was built from $N = 1{,}000$ repetitions. Red lines represent identity functions. Black continuous lines (for simulations drawn with $T \sim$ U[-10,000-0]) with black triangles and gray dashed lines ($T \sim$ U[-30,000-0]) with gray circles indicate means for the corresponding estimated values (y-axis) over small intervals for the real values (x-axis). Gray triangles (for simulations drawn with $T \sim$ U[-10,000-0]) and gray circles ($T \sim$ U[-30,000-0]) in the background represent all 1,000 values obtained for each cross-validation model. (C) Confusion matrix for 1,000 randomly chosen simulated variants drawn as in the right panel of (B) with a prior distribution of $T \sim$ U[-10,000-0], 500 with an onset of selection <4,500 ya and 500 with an onset of selection >4,500 ya (left panel), or for 1,000 randomly chosen simulated variants drawn as in the right panel of (B) with a prior distribution of $T \sim$ U[-30,000-0], 500 with an onset of selection <4,500 ya and 500 with an onset of selection >4,500 ya (right panel). The goodness-of-fit for the method is summarized as an F1-score.

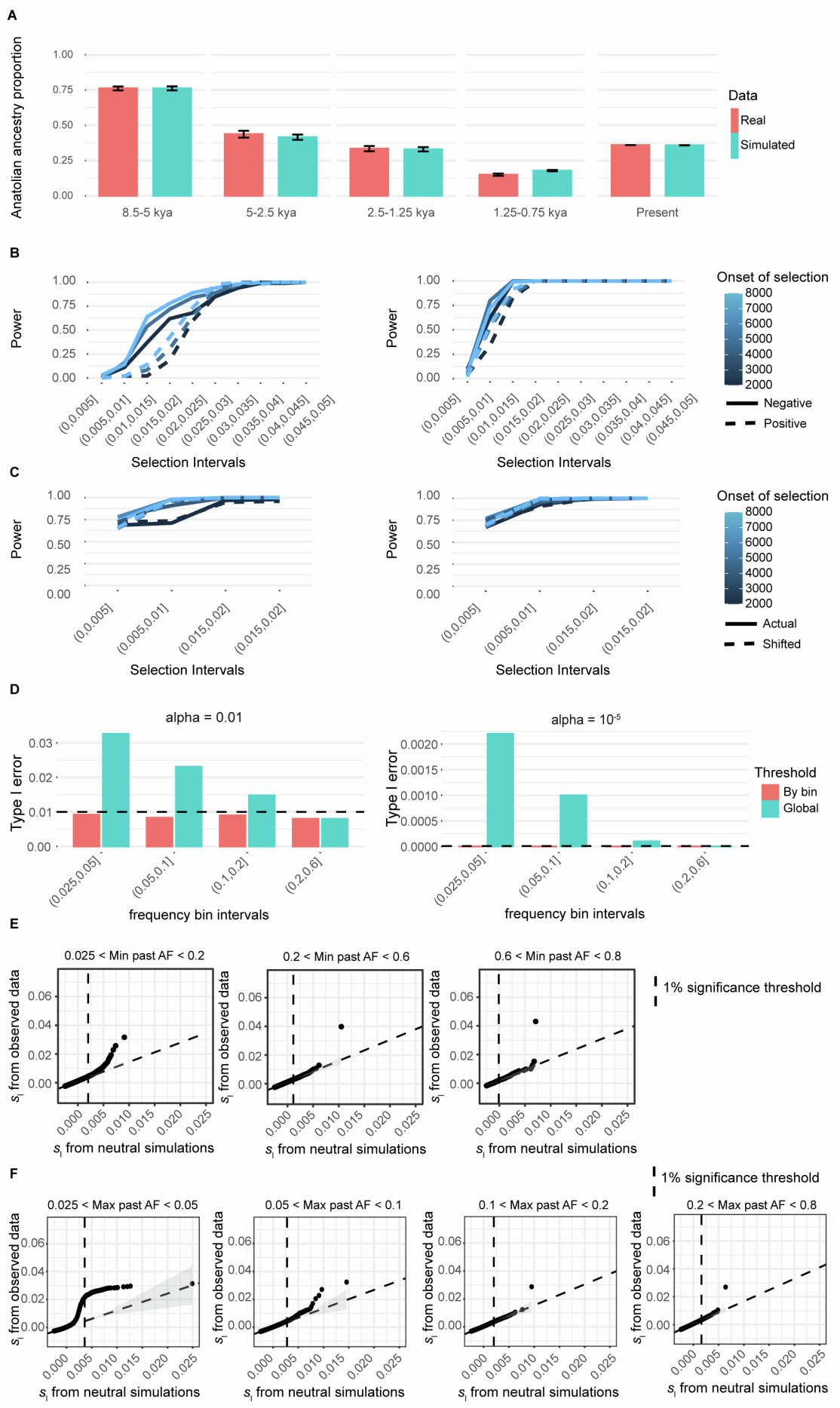
A
Anatolian ancestry proportion
1.00
0.75
0.50
0.25
0.00
8.5-5 kya
5-2.5 kya
2.5-1.25 kya
1.25-0.75 kya
Present
Data
Real
Simulated
B
Power
Selection Intervals
Onset of selection
8000
7000
6000
5000
4000
3000
2000
Negative
Positive
C
Actual
Shifted
D
alpha = 0.01
alpha = 10-5
Type I error
frequency bin intervals
Threshold
By bin
Global
E
0.025 < Min past AF < 0.2
0.2 < Min past AF < 0.6
0.6 < Min past AF < 0.8
si from observed data
si from neutral simulations
1% significance threshold
F
0.025 < Max past AF < 0.05
0.05 < Max past AF < 0.1
0.1 < Max past AF < 0.2
0.2 < Max past AF < 0.8

Figure S2. Analyses relating to selection detection, Related to STAR Methods
(A) Mean Anatolian ancestry proportion estimated across all individuals from a given period (x-axis), either for empirical data (red) or simulated data (blue; using ABC simulations as described in **STAR Methods**). Standard deviations around the mean are shown with the error bar on top of the plotted bars.
(B) Power to detect positive selection (dashed lines) and negative selection (continuous lines), for an allele frequency distribution matching real data for variants with frequencies between 0 and 0.2 (left panel) or between 0.2 and 0.6 (right panel) for an empirical significance level of 10^{-5}. Power estimates for each combination of selection coefficient and time of selection onset are based on 100 selection coefficient estimations.
(C) Power to detect positive selection for an empirical significance level of 10^{-2} using the frequency bin boundaries used in this work (continuous lines; **STAR Methods)**, i.e., [0.025-0.2], [0.2-0.6] and [0.6-0.8]. Dashed lines indicate, on the contrary, the use of shifted frequency bin boundaries, here [0.025-0.1], [0.1-0.5] and [0.5-0.8].
(D) Type I error estimations for a 1% (left panel) or a 10^{-5} (right panel) nominal level of significance for either significance thresholds of s_l obtained within each frequency bin (red) or a global and unique significance threshold used for all variant categories (blue).
(E) QQ-plots for the distributions of s_l derived from neutral simulations (x-axis) or observed data (y-axis) within each tested frequency bin, for the analysis of positive selection. Vertical dashed lines indicate the 1% significance threshold in each case. 'Min past AF' indicates the minimum, across past epochs, of the upper bound of the allele frequency's 95% CI.
(F) Distributions of s_l as in (E) but for the analysis of negative selection. 'Max past AF' indicates the maximum, across past epochs, of the lower bound of the allele frequency's 95% CI across all past epochs.

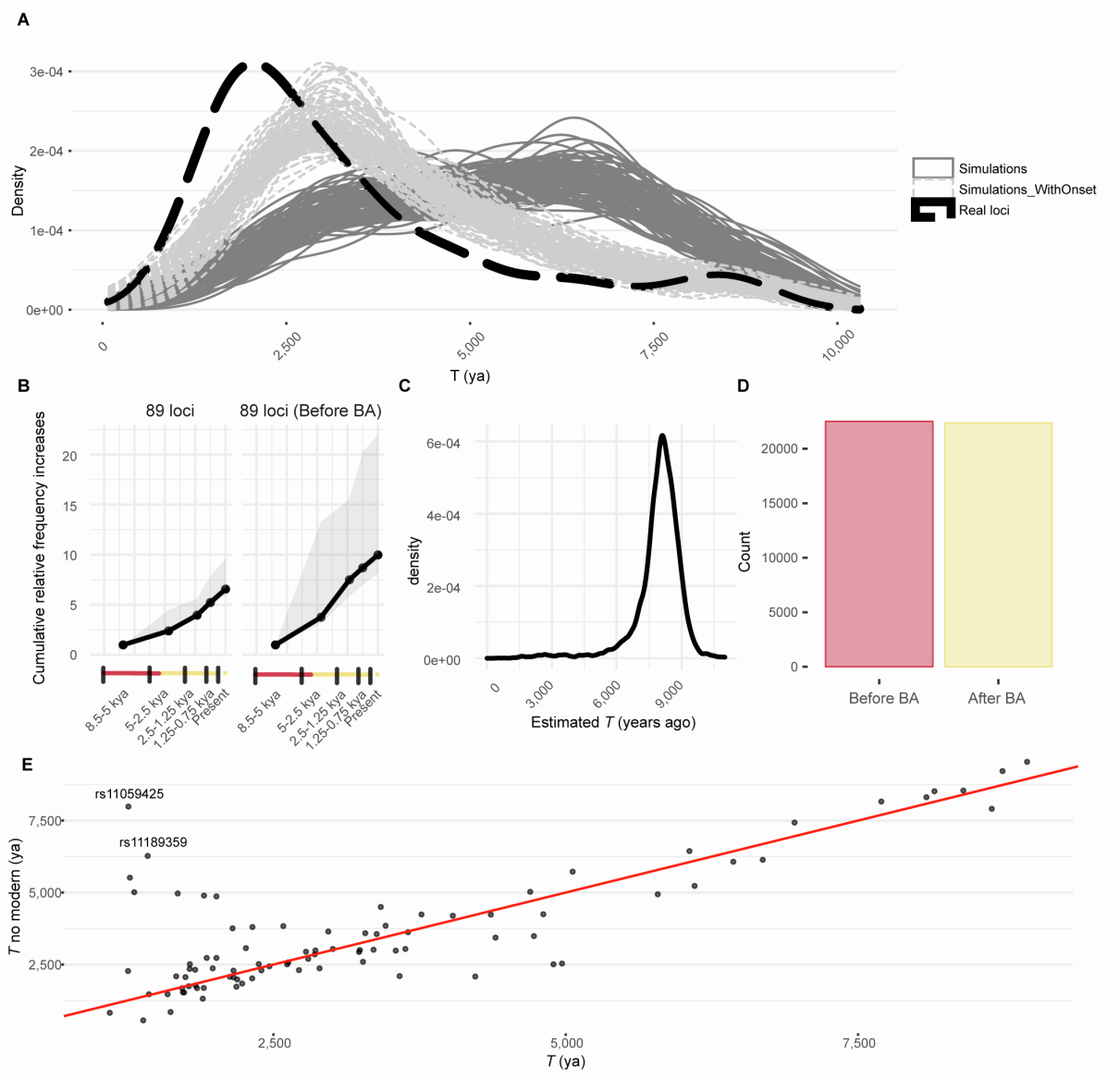

Figure S3. Time of onset of selection for candidate positively selected loci, Related to Figure 1 and STAR Methods

(A) Density distributions of selection onset estimates (T) for the variant with the smallest p_{sel} at each of the 89 positively selected loci (dashed black line), for 100 sets of 89 independent simulated variants matching the allele frequencies and selection strengths of each of the 89 variants (continuous dark gray lines), or for 100 sets of 89 independent simulated variants matching the allele frequencies, selection strengths and times of selection onset of the 89 variants (continuous light gray lines).

(B) Increases in frequency relative to the previous period were computed for the top 89 variants (**Table S2**) (left panel) or for those with estimated onsets of selection before the start of the Bronze Age (right panel). The average at each period was computed and the cumulative values across epochs were plotted as black dots. This was also conducted for the lower and the upper bound of the 95% CI of the corresponding allele frequencies and their values indicate the limits of the light gray surface.

(C) Density plot for the distribution of estimated T values from N = 30,000 simulated variants under old, short-lived positive selection. Under this scenario, we assumed that $s \sim U[0.01,$

0.05] and selection starts at $T \sim U[6000, 10000]$ ya, and stops at max(0, $T + 3000$) ya. The evolutionary parameters were randomly sampled from uniform distributions and the estimations of T were derived from the ABC approach assuming constant-in-time selection, as done throughout this work.
(D) Number of detected events of selection for which T was estimated before (red) or after (yellow) the start of the Bronze Age, from $N = 80{,}000$ simulated variants under short-lived positive selection. Under this scenario, we assumed that $s \sim U[0.01, 0.05]$ and selection starts at $T \sim U[6000, 10000]$ ya, and stops at max(0, $T + 3000$) ya. As in (B) the evolutionary parameters were randomly sampled from uniform distributions.
(E) T estimates for the 89 positively selected variants with minimal p values at each of the 89 loci determined with (x-axis) or without (y-axis) using modern DNA data.

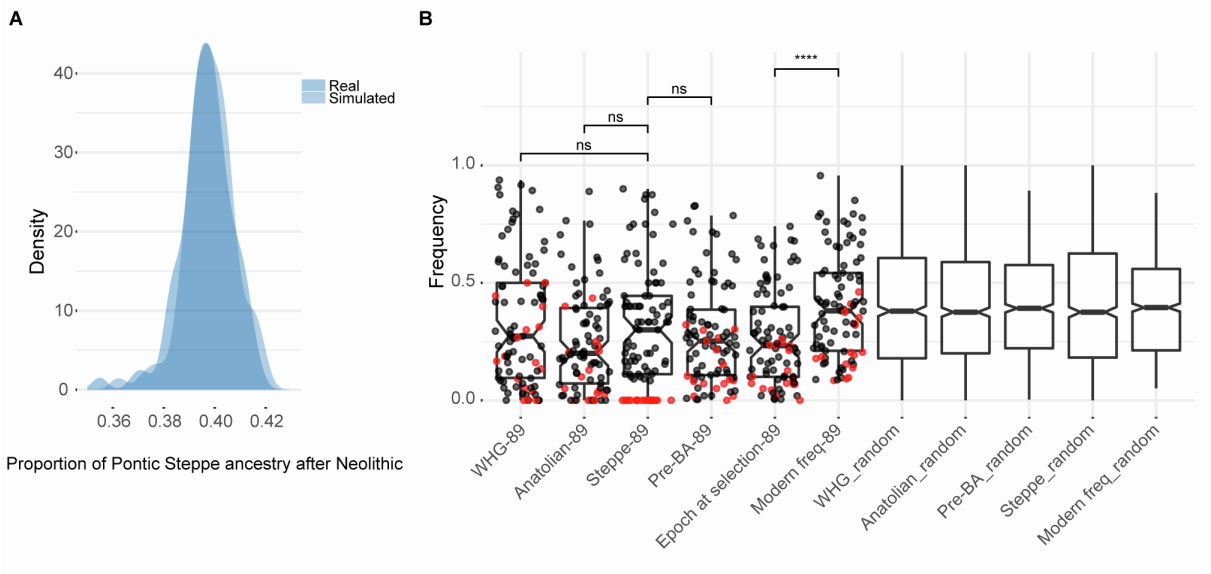

Figure S4. Adaptive admixture has not been the main driver of positive selection in post-Neolithic Europe, Related to Figure 1

(A) Distribution of the estimated mean proportion of Pontic Steppe ancestry for the post-Neolithic carriers of the allele with smallest p_{sel} at each of the 89 candidate positively selected loci (darker blue, 'Real'), or for the simulated aDNA used to obtain s and T estimates for each of the 89 variants (lighter blue, 'Simulated').

(B) Frequencies for the 89 alleles with smallest p_{sel} at each of the 89 candidate positively selected loci (boxplots with dots) or for 1,000 random variants matched on DAF (boxplots without dots). Frequency distributions are represented for different ancestry groups (Western hunter gatherer [WHG], Anatolian and Steppe) and differences between these distributions were assessed with a Wilcoxon test (not significant [ns] for all three tests). Frequencies are represented for different epochs (pre-Bronze Age samples [Pre-BA], samples from the epoch in which selection is estimated to have begun [Epoch at selection] and modern samples [Modern Freq]). A Wilcoxon test was also used to test for differences in frequency distributions between "Epoch at selection-89" and "Modern Freq-89". This difference was significant, as expected. The red dots indicate alleles with a frequency estimation of 0 among individuals with Pontic-Steppe ancestry. Individuals were defined as belonging to a given ancestry category if their proportion of the corresponding ancestry was >90%. The Steppe category was also limited to individuals dating back more than 4,500 years, the Anatolian category was limited to individuals dating back more than 7,500 years, and the WHG category was limited to individuals dating back more than 8,000 years. Frequency trajectories for these 89 alleles can be found in **Table S2**.

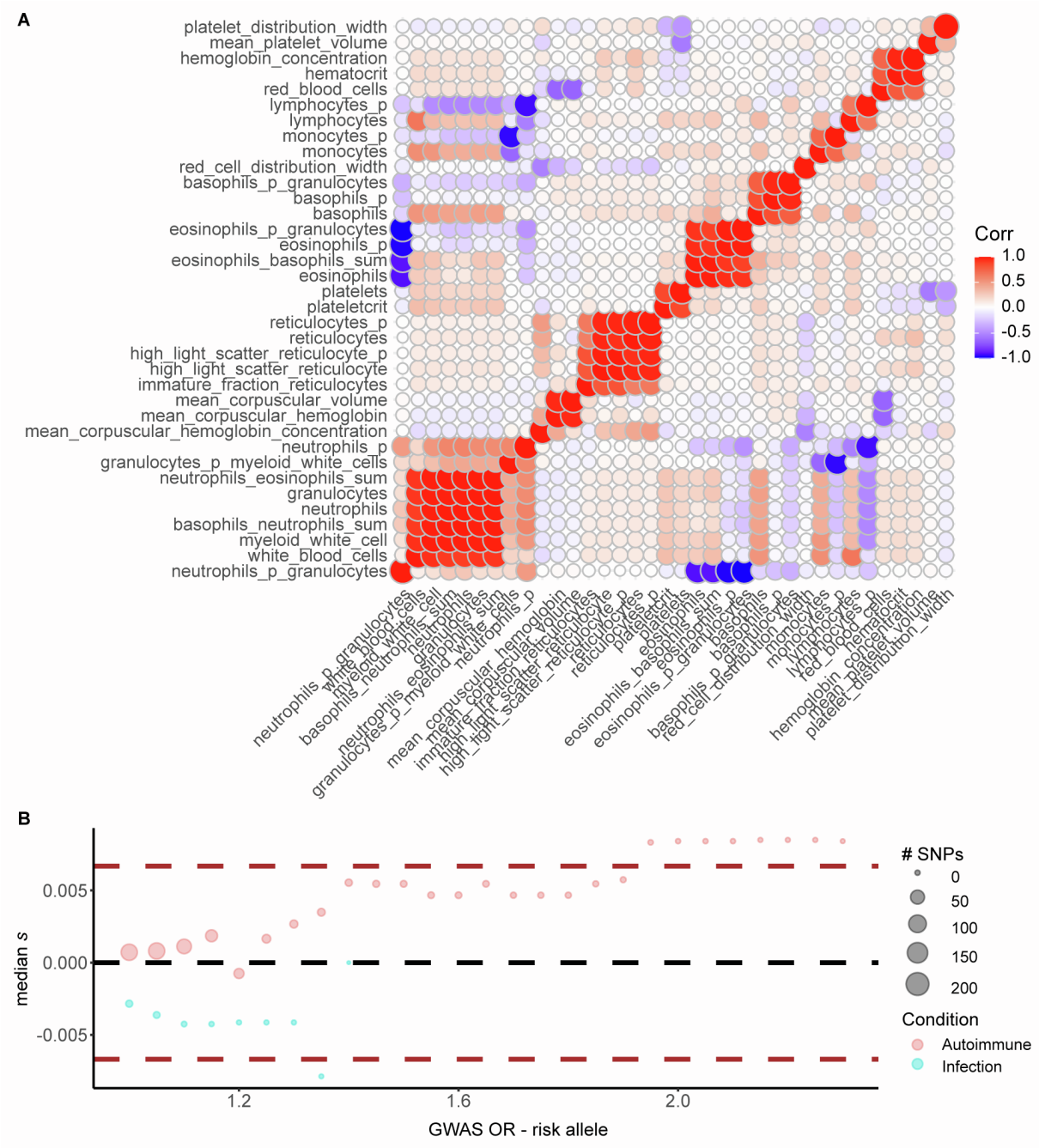

Figure S5. Genetic correlations between hematopoietic traits and relationship between odds-ratio and *s* at variants associated with autoimmune and infectious diseases, Related to Figure 2 and Figure 3

(A) Genetic correlations were computed using 'ldsc' (**STAR Methods**).

(B) Each dot represents the median *s* coefficient (y-axis) for the subset of variants contributing to the PRS of infectious (blue) or autoimmune (red) disorders with estimated GWAS odds-ratio (OR) indicated in the x-axis. Dots are drawn for subsets consisting of at least two variants. Dot size scales with the numbers of variants in the subset.

Table S1. Evolutionary parameters for previously published positively selected loci, Related to Figure 1

CHR	Lead SNP[a]	POS hg19	DA	S^b	T (ya)[b]	Genes[c]	p_{sel}
2	*rs4988235*	135.3-137.3	A	0.082 (0.087)	6,102 (3,401)	*MCM6, LCT*	0
5	*rs185146*	33.8-34	T	0.059 (0.061)	4,226 (12,583)	*SLC45A2*	0
6	*rs3130673*	29.9-33.1	T	0.077 (0.067)	8,401 (10,901)	*MHC region*	0
11	*rs174537*	61.5-61.6	G	0.013 (0.013)	6,685 (10,263)	*FADS1, FADS2*	2.7×10^{-5}
4	*rs13149231*	38.7-38.8	T	0.016 (0.020)	6,570 (11,300)	*TLR1*	5.8×10^{-5}
12	*rs11065987*	111.9-112.6	G	0.013 (0.028)	5,597 (6,456)	*ATXN2, SH2B3*	5.2×10^{-4}
11	*rs1540129*	71.1-71.2	C	0.020 (0.021)	3,333 (3,251)	*DHCR7*	0
11	*rs10765770*	88.5-88.9	A	0.019 (0.021)	6,058 (8,618)	*GRM5*	2.0×10^{-6}
5	*rs27879*	131.4-131.8	A	0.020 (0.027)	8,884 (11,957)	*SLC22A4*	1.9×10^{-5}
6	*rs9468413*	28.3-28.7	C	0.049 (0.042)	6,057 (8,927)	*ZKSCAN3, ZSCAN31*	0
13	*rs9566230*	38.1-38.8	C	0.017 (0.029)	1,530 (26,076)	-	3.8×10^{-5}
11	*rs72878978*	29.9-30.1	A	0.016 (0.011)	2,387 (1,865)	*HERC2, OCA2*	2.9×10^{-3}

[a]The SNP with the smallest p_{sel} at each of the 12 previously published positively selected loci.[S1]
[b]T and s values are shown for estimates based on $T \sim$ U[-10,000-0] ya or in parentheses for $T \sim$ U[-30,000-0] ya, as priors used in the simulations.
[c]Most relevant gene(s) for the lead SNP of each locus.

Table S3. Variants underlying the evolutionary history of the ABO locus, Related to Figure 1

SNP (ABO type)	In 1240k array	Proxy (allele)	Correlated alleles - r^2 (ABO type)	p_{sel} (negative or positive)	Disease association (direction)
rs8176719>TC (non-O type)	No	No	-	-	Malaria (susceptibility)
rs8176746>G (B type)	No	No	-	-	-
rs41302905>T (O type)	Yes	No	-	0.04 (negative)	-
rs505922>C	Yes	Yes (rs8176719)	C=TC - 0.87 (non-O type)	0.01 (positive)	-
rs8176749>T	Yes	Yes (rs8176746)	T=G - 1 (*B* type)	1.2×10^{-4} (positive)	-
rs8176743>T	Yes	Yes (rs8176746)	T=G - 1 (B type)	0.05 (positive)	-
rs9411378>A	No	Yes (rs8176719)	A=TC - 0.64	-	COVID-19 (susceptibility)
rs635634>C	Yes	Yes (rs8176719)	C=TC - 0.4	6.4×10^{-4} (positive)	TS and CIE (protection)

Note: TS stands for tonsillectomy and CIE for childhood ear infection. r^2 are given for EUR populations of 1KG. We found no variant correlated with rs9411378 with an $r^2 > 0.8$ in the 1240k array.

Table S5. Top SNPs underlying the increase in the risk of inflammatory disorders over time, Related to Figure 3

SNP	Phenotype	Consequence	*p* (PRS)[a]	*p* (eQTL)	Gene[b]
rs2188962	CD	Intronic	7.04×10^{-19}	1.8×10^{-20}	*IRF1*
rs2188962	IBD	Intronic	5.09×10^{-19}	1.8×10^{-20}	*IRF1*
rs11066188	CD	Upstream	2.82×10^{-12}	1.2×10^{-49}	*SH2B3*
rs11066188	IBD	Upstream	1.80×10^{-11}	1.2×10^{-49}	*SH2B3*
rs492602	CD	Synonymous	2.08×10^{-10}	-	*FUT2*
rs1456896	IBD	Intergenic	2.25×10^{-4}	1.6×10^{-3}	*IKZF1*
rs10774679	COVID	Upstream	2.05×10^{-4}	-	*OAS1*

[a]*p* values were obtained after adjustment for ancestry (factor components) and sample location (longitude and latitude)
[b]Either the gene eQTL/sQTL (rs2188962, rs11066188, rs1456896 and rs10774679) or the gene in which the variant occurs (rs492602)

Table S6. Enrichment of missense and low-DAF variants among candidate negatively selected variants, Related to Figure 4

SNP subset	$p_{sel} < 0.01$	$p_{sel} < 10^{-4}$
missense	OR = 1.1 95% CI = [1-1.3]	OR = 1.8 95% CI = [1.2-2.7]
intronic	OR = 1 95% CI = [1-1.1]	OR = 1 95% CI = [0.8-1.1]
0.025 < Max past AF <0.05	OR = 4.8 95% CI = [4.3-5.4]	OR = 198 95% CI = [85-648]
0.05 < Max past AF <0.1	OR = 2.4 95% CI = [2.2-2.6]	OR = 14 95% CI = [7.3-31]
0.1 < Max past AF <0.2	OR = 3.1 95% CI = [2.9-3.3]	OR = 6.1 95% CI = [3.5-12]
0.2 < Max past AF <0.6	OR = 4.2 95% CI = [4.1-4.4]	OR = 29 95% CI = [21-40]

Note: 'Max past AF' indicates the maximum, across past epochs, of the lower bound of the allele frequency's 95% CI.

Table S9. Filtering bias for variants close to indels or not in the capture dataset, Related to STAR Methods

SNP	Shotgun_c[a] Capture_na	Shotgun_nc[a] Capture_na	Capture[b] frequency	Shotgun[b] frequency	1KG frequency	Nearby indel
rs264272	0.72	0.014	0.12	0.49	0.51	Yes
rs2792600	0.78	0.051	0.05	0.38	0.45	Yes
rs4283567	0.54	0.027	0.19	0.41	0.49	Yes
rs8076492	0.61	0.054	0.28	0.56	0.56	Yes
rs13225222	0.57	0.025	0.04	0.14	0.19	Yes
rs7248807	0.73	0.022	0.07	0.26	0.39	Yes
rs11621931	0.67	0.024	0.01	0.08	0.10	Yes
rs13126794	0.71	0.030	0.02	0.15	0.17	Yes
rs10763588	0.59	0.046	0.20	0.54	0.56	Yes
rs996759	0.041	0.026	0.42	0.28	0.45	No
rs13180583	0.052	0.029	0.40	0.37	0.39	No
rs6437191	0.079	0.017	0.40	0.44	0.41	No
rs16961871	0.034	0.029	0.13	0.15	0.20	No
rs12615153	0.026	0.025	0.56	0.60	0.67	No
rs2835900	0.037	0.034	0.27	0.29	0.35	No
rs10146186	0	0.024	0.07	0.06	0.06	No
rs11592504	0.11	0.094	0.07	0.05	0.10	No
rs72649559	0.018	0.026	0.16	0.14	0.19	No

Note: Columns 2 and 3 indicate the proportion of individuals with a status in the shotgun database but absent from the capture database.

[a]'_c' stands for carriers, '_nc' stands for non-carrier and '_na' stands for absent

[b]Frequencies were obtained by averaging frequencies across all past epochs

Table S10. Demographic parameters used for the simulation study, Related to STAR Methods

Parameter	Minimum value	Maximum value
***Split* AFR-NonAFR (*kya*)**	60	80
***Ne* (NonAFR)**	2,000	3,000
***Split* WestNonAFR-EastNonAFR (*kya*)**	35	45
***Ne* (WestNonAFR)**	3,000	4,000
***Split* EUR-Anatolia/Steppe (*kya*)**	15	25
***Ne* (Anatolia/Steppe)**	= Ne WestNonAFR	= Ne WestNonAFR
***Pulse*-Anatolia (*kya*)**	8	10
***Pulse*-Steppe (*kya*)**	4	6
***Intensity Pulse*-Anatolia**	*	*
***Intensity Pulse*-Steppe**	*	*

Note: Minimum and maximum values define the range for each parameter. A uniform distribution was used to draw values from these ranges. *Split* stands for divergence time, *Ne* for effective population size, *Pulse* for instantaneous migration (accounting for *x*% of migrants of the source population, where *x* is given by the intensity of the pulse). Populations follow an exponential growth over the last 25,000 years following the formula established by Gravel et al.[S2] i.e., $initial\_pop\_size \times (1 + growth\_coef)^t$, where for example *growth_coef* = 0.0038 for the EUR population, *initial_pop_size* is the initial population size 25,000 ya, and *t* increases as the generations since the start of the exponential growth.
*Intensity of pulses are randomly sampled from a normal distribution with means and standard deviations estimated from the computed ancestry proportion of aDNA samples with the 50% largest genome coverage and belonging to the relevant epoch (Neolithic for the Anatolian pulse and Bronze Age for the Steppe pulse).

Table S11. Primers used for experiments on *LBP*, Related to Figure 5 and STAR Methods

Name		Sequence		Used for
LBP_cDNA1_F	5'	ACTGCTGGCATTGCTGCTT	3'	Sequencing of LBP cDNA
LBP_cDNA1_R	5'	ACTTGGACTCAATCTGGTTGTGG	3'	
LBP_cDNA2_F	5'	CAGTTACTGCCTCCAGCTGCA	3'	
LBP_cDNA2_R	5'	GCACTGATCCCTGGAGTTCCAG	3'	
LBP_cDNA3_F	5'	ATTATGTCTTCAACACGGCCAGC	3'	
LBP_cDNA3_R	5'	CAGGAAGTCCTTATGGATCTGCAG	3'	
LBP_cDNA4_F	5'	GATCACTGGGTTCCTGAAGCCAG	3'	
LBP_D283G_F	5'	TTGCCATCTCGGGTTATGTCTTCAACAC	3'	Targeted mutagenesis
LBP_D283G_R	5'	GTGTTGAAGACATAACCCGAGATGGCAA	3'	
LBP_P333L_F	5'	CCAGGCTCTACCTCAACATGAACC	3'	
LBP_P333L_R	5'	GGTTCATGTTGAGGTAGAGCCTGG	3'	
LBP_stop_F	5'	ATACATGAGAGTTTAGCGTACGCGGC	3'	Deletion of C-terminal DDK tag
LBP_stop_R	5'	GCCGCGTACGCTAAACTCTCATGTAT	3'	

SUPPLEMENTAL REFERENCES

S1. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499-503. 10.1038/nature16152.

S2. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Project, T.G., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. PNAS *108*, 11983-11988. 10.1073/pnas.1019276108.