Supplementary Material for

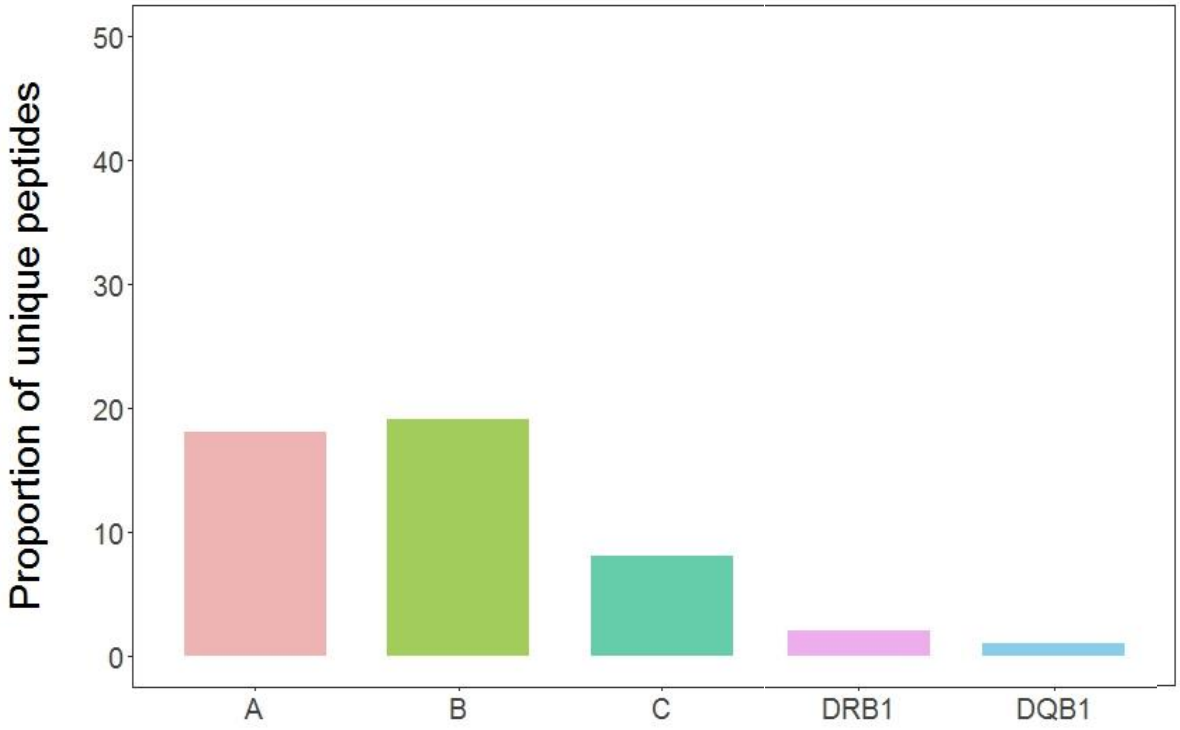
Divergent allele advantage at human MHC genes: signatures of past and ongoing selection

Federica Pierini [1] & Tobias L. Lenz [1,*]

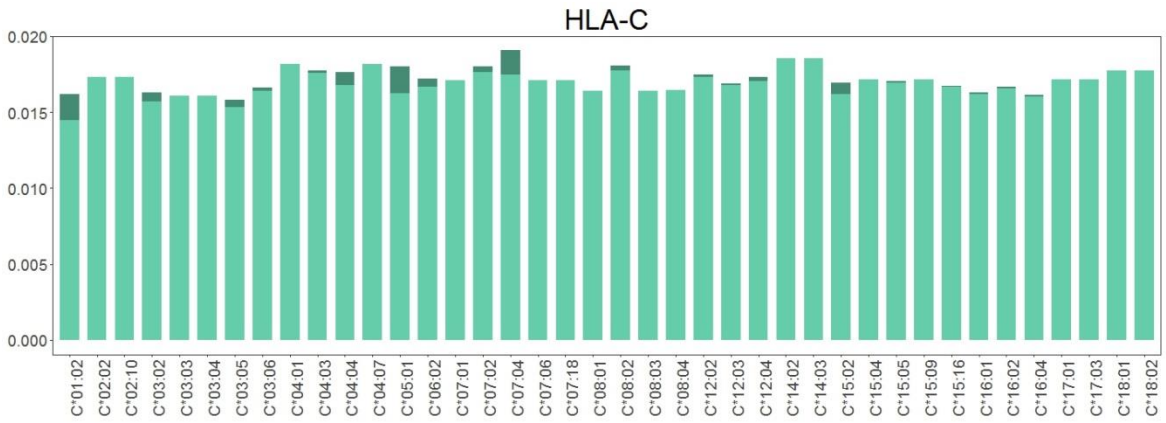
* Corresponding author: Tobias L. Lenz, lenz@post.harvard.edu

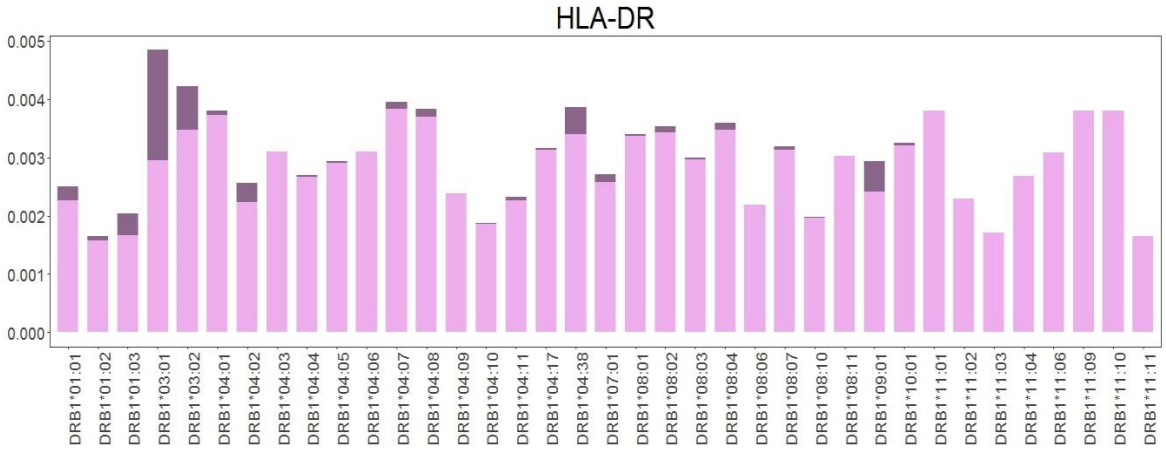
Supplementary material includes:
Supplementary Figures S1-S11
Supplementary Tables S1-S9

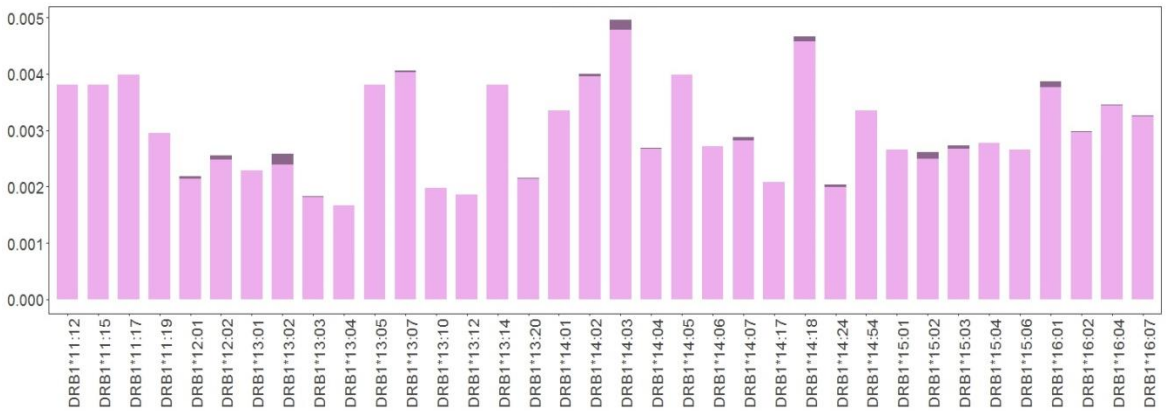
Supplementary figure S1. Proportion of unique peptides (out of a total of 118,097 pathogen-derived peptides) predicted to be bound by at least one allele is shown for each locus.

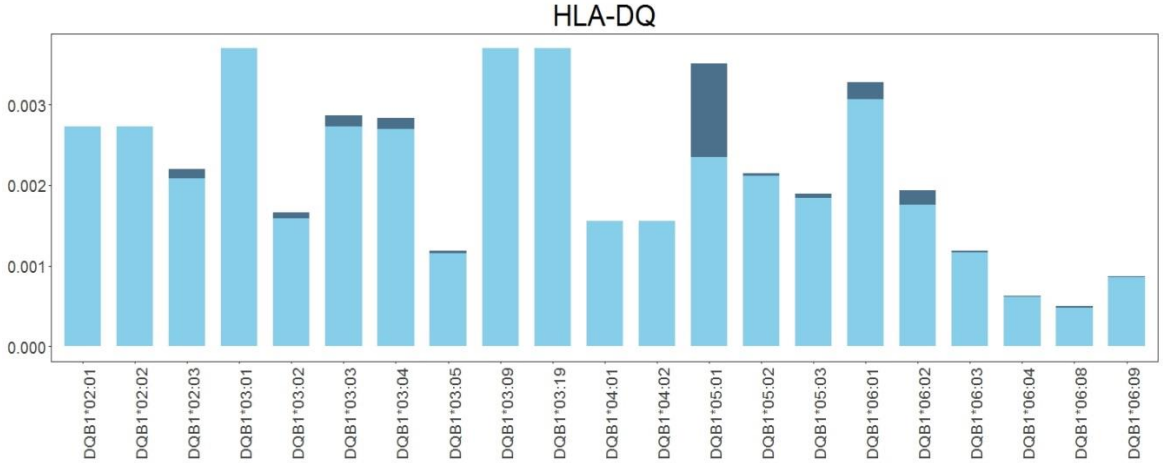


Supplementary figure S2. Percentage of common (light colours) and private (dark colours) peptides predicted to be bound by each allele at the different MHC loci.

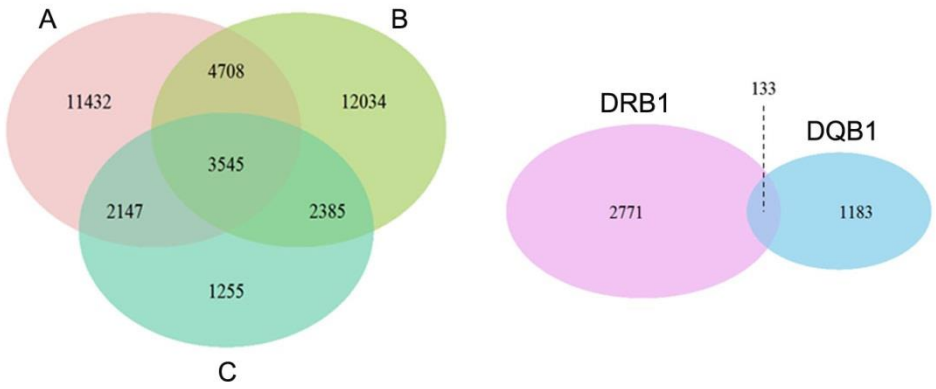

HLA-C
0.020
0.015
0.010
0.005
0.000
C*01:02
C*02:02
C*02:10
C*03:02
C*03:03
C*03:04
C*03:05
C*03:06
C*04:01
C*04:03
C*04:04
C*04:07
C*05:01
C*06:02
C*07:01
C*07:02
C*07:04
C*07:06
C*07:18
C*08:01
C*08:02
C*08:03
C*08:04
C*12:02
C*12:03
C*12:04
C*14:02
C*14:03
C*15:02
C*15:04
C*15:05
C*15:09
C*15:16
C*16:01
C*16:02
C*16:04
C*17:01
C*17:03
C*18:01
C*18:02


HLA-DR
0.005
0.004
0.003
0.002
0.001
0.000
DRB1*01:01
DRB1*01:02
DRB1*01:03
DRB1*03:01
DRB1*03:02
DRB1*04:01
DRB1*04:02
DRB1*04:03
DRB1*04:04
DRB1*04:05
DRB1*04:06
DRB1*04:07
DRB1*04:08
DRB1*04:09
DRB1*04:10
DRB1*04:11
DRB1*04:17
DRB1*04:38
DRB1*07:01
DRB1*08:01
DRB1*08:02
DRB1*08:03
DRB1*08:04
DRB1*08:06
DRB1*08:07
DRB1*08:10
DRB1*08:11
DRB1*09:01
DRB1*10:01
DRB1*11:01
DRB1*11:02
DRB1*11:03
DRB1*11:04
DRB1*11:06
DRB1*11:09
DRB1*11:10
DRB1*11:11

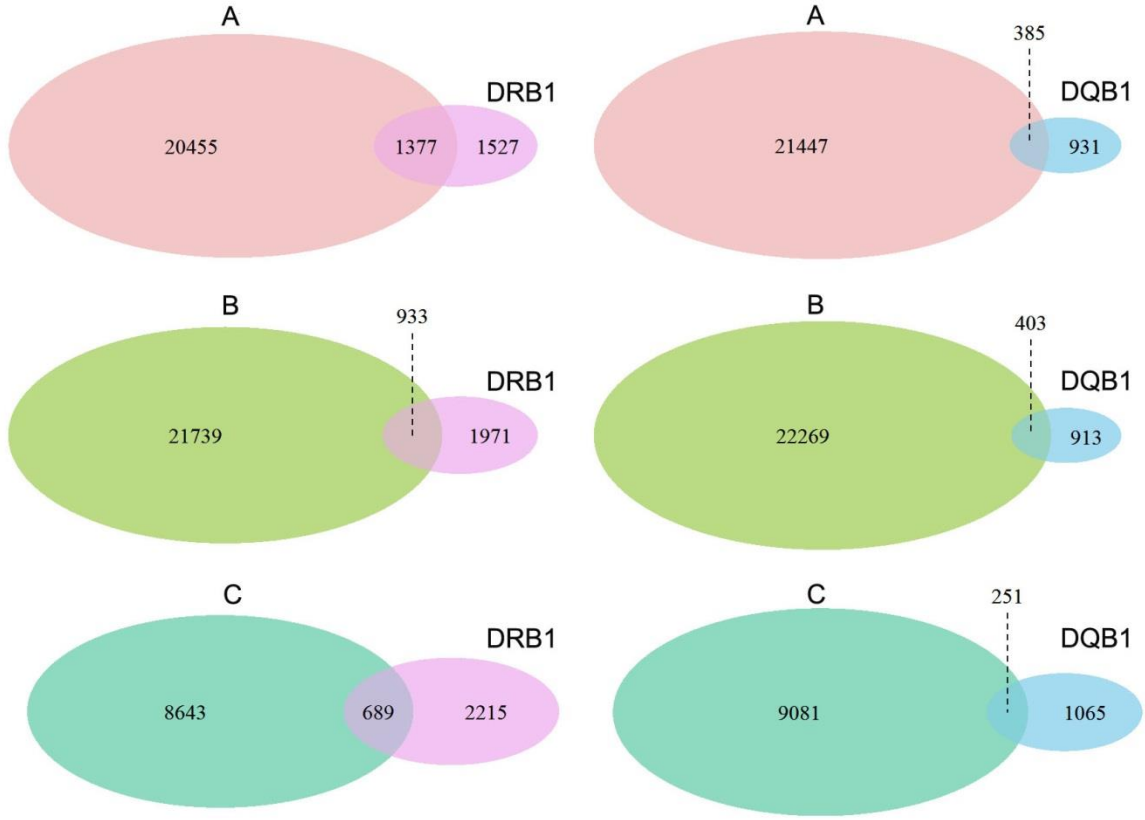

0.005
0.004
0.003
0.002
0.001
0.000
DRB1*11:12
DRB1*11:15
DRB1*11:17
DRB1*11:19
DRB1*12:01
DRB1*12:02
DRB1*13:01
DRB1*13:02
DRB1*13:03
DRB1*13:04
DRB1*13:05
DRB1*13:07
DRB1*13:10
DRB1*13:12
DRB1*13:14
DRB1*13:20
DRB1*14:01
DRB1*14:02
DRB1*14:03
DRB1*14:04
DRB1*14:05
DRB1*14:06
DRB1*14:07
DRB1*14:17
DRB1*14:18
DRB1*14:24
DRB1*14:54
DRB1*15:01
DRB1*15:02
DRB1*15:03
DRB1*15:04
DRB1*15:06
DRB1*16:01
DRB1*16:02
DRB1*16:04
DRB1*16:07


HLA-DQ
0.003
0.002
0.001
0.000
DQB1*02:01
DQB1*02:02
DQB1*02:03
DQB1*03:01
DQB1*03:02
DQB1*03:03
DQB1*03:04
DQB1*03:05
DQB1*03:09
DQB1*03:19
DQB1*04:01
DQB1*04:02
DQB1*05:01
DQB1*05:02
DQB1*05:03
DQB1*06:01
DQB1*06:02
DQB1*06:03
DQB1*06:04
DQB1*06:08
DQB1*06:09

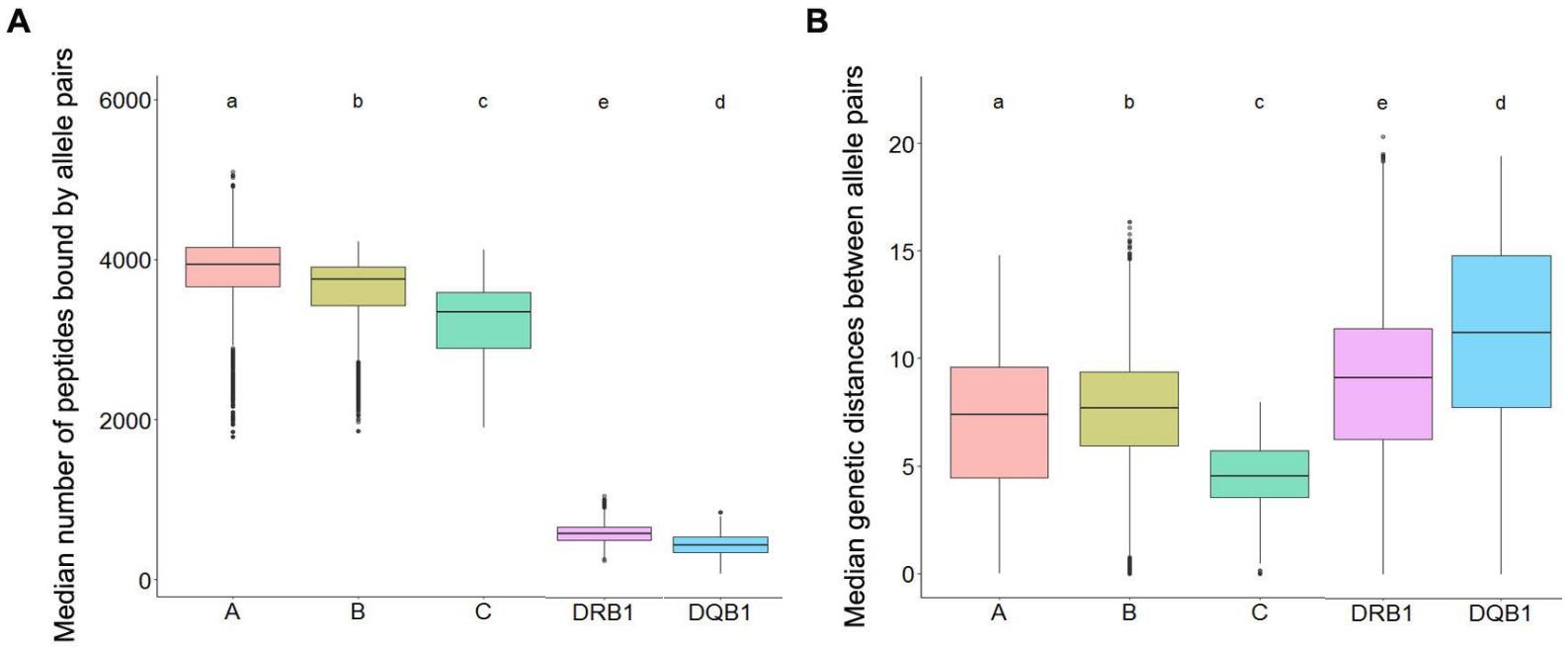
Supplementary figure S3. Overlap in bound peptides between different loci for alleles at class I (A ∩ B ∩ C) and class II (DQB1 ∩ DRB1) genes. Number of common alleles tested at the different HLA loci: HLA-A: 63, HLA-B: 123, HLA-C: 40, HLA-DRB1: 73, HLA-DQB1: 21. Binding prediction analyses performed on the complete dataset of pathogen proteins (n=232) for a total of 118,097 unique pathogen-derived peptides. Bound peptides represent the total number of unique peptides bound by at least one allele at each locus.



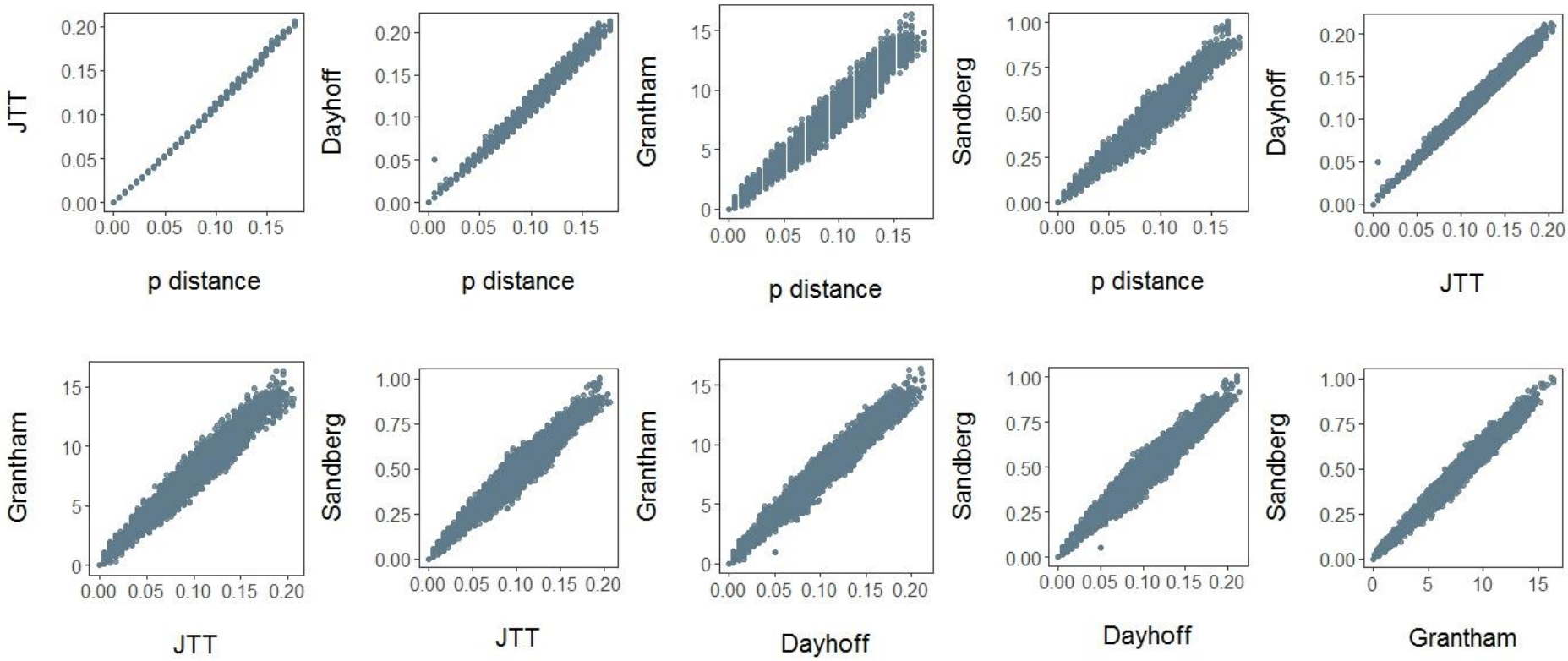
Supplementary figure S4. Overlap in bound peptides between class I and class II genes (A ∩ DRB1, A ∩ DQB1, B ∩ DRB1, B ∩ DQB1, C ∩ DRB1, C ∩ DQB1).



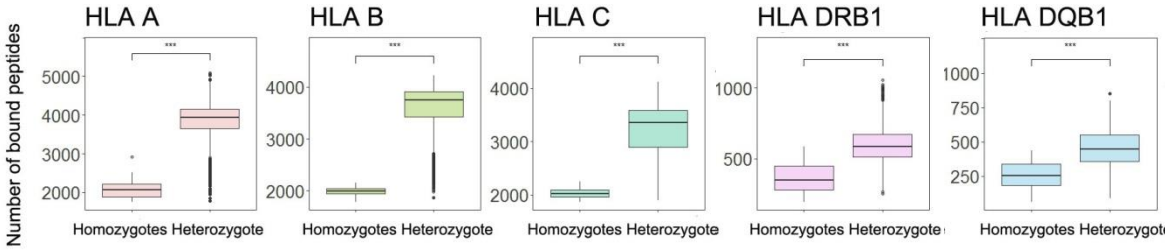
Supplementary figure S5. **(a)** Median number of peptides bound by allele pairs and **(b)** median genetic distances between allele pairs at each locus. Comparison of groups was done using the Kruskal-Wallis test followed by multiple comparison tests for pairwise differences. Different letters indicate statistically significant differences between loci ($p < 0.001$).



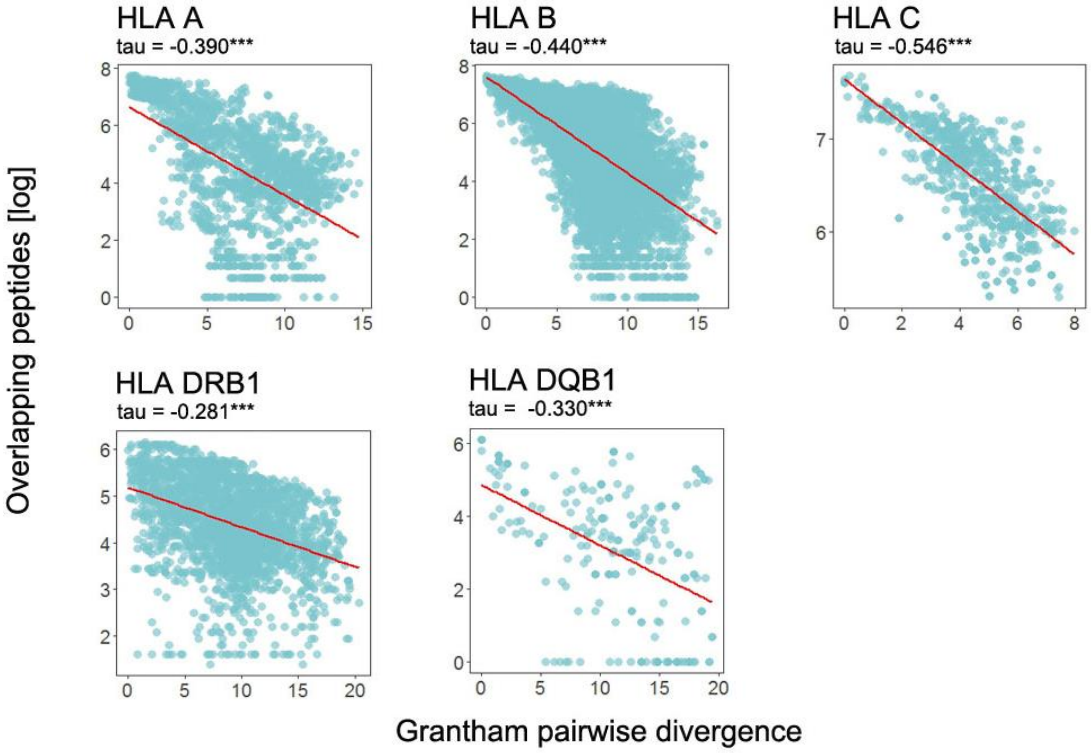
Supplementary figure S6. Correlation between pairwise parameters of allele divergence for HLA-B locus.



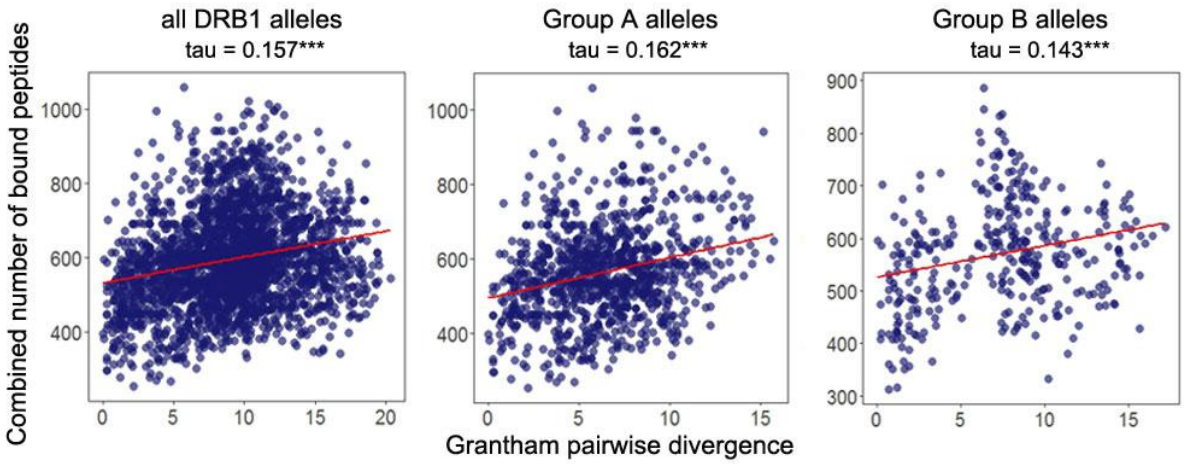
Supplementary figure S7. Heterozygote advantage. Total number of combined peptides bound by all possible homozygote and heterozygote genotypes at each locus (HLA-A, homozygotes n = 62, heterozygote n = 1891; HLA-B, homozygotes n = 123, heterozygote n = 7503; HLA-C, homozygotes n = 40, heterozygote n = 780; HLA-DRB1, homozygotes n = 73, heterozygote n = 2628; HLA-DQB1, homozygotes n = 21, heterozygote n = 210).



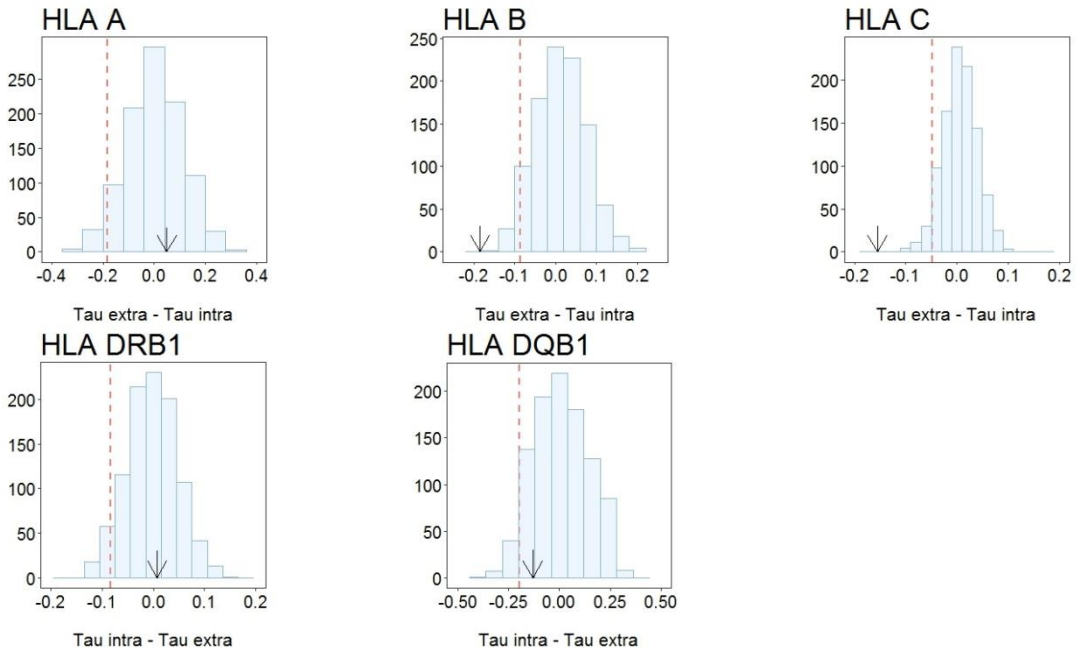
Supplementary figure S8. Correlation between pairwise genetic distances reported as Grantham distance (x-axes) and the number of overlapping peptides reported on a logarithmic scale (y-axes) counted for all possible pairs of common HLA alleles. Each dot represents an allele pair. Binding prediction analyses performed on the complete dataset of pathogen proteins (n=232).



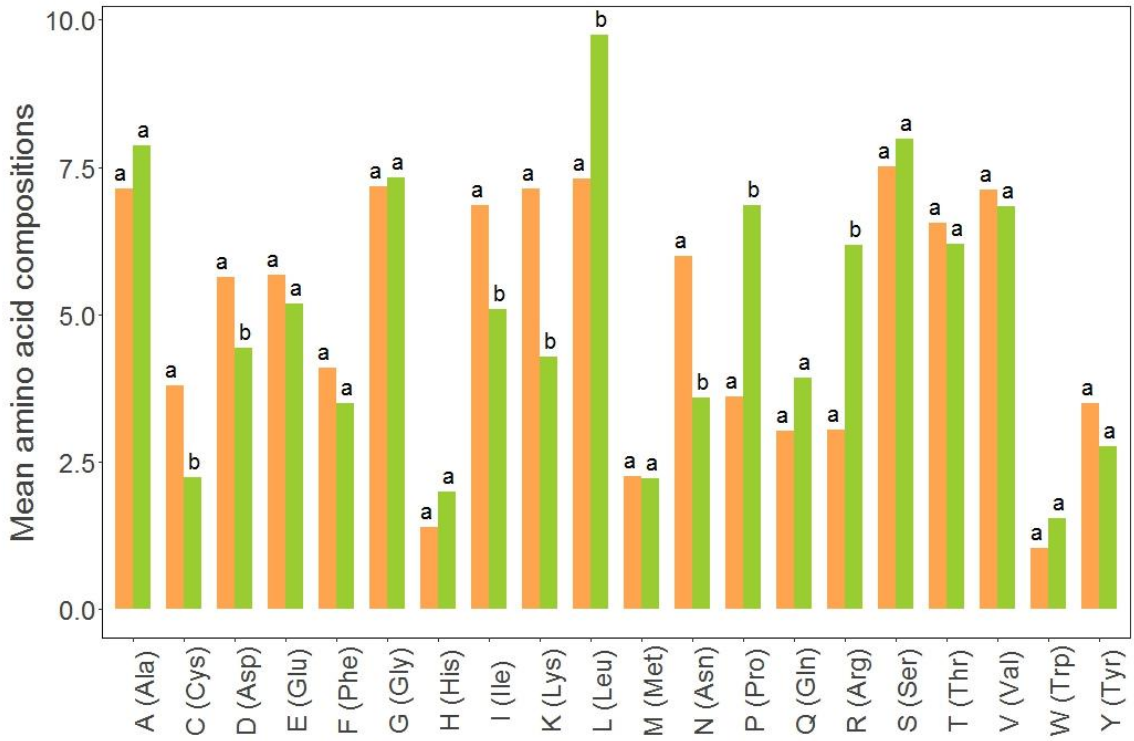
Supplementary figure S9. Correlation between pairwise genetic distances reported as Grantham distance (x-axes) and the number of bound peptide (y-axes) counted for all possible allele pairs. Each dot represents an allele pair. Binding prediction analyses performed on the complete dataset of pathogen proteins (n=232) and considering all common DRB1 alleles as well as Group A (N = 46) and Group B (N = 27) DRB1 alleles individually.



Supplementary figure S10. Distribution of differences in correlations values between intracellular and extracellular proteins (light blue histograms) obtained by permuting original proteins 1000 times between the three groups of pathogens and keeping the same number of proteins for each group as it occurred in the initial dataset. For each permuted dataset, the coefficient (Kendall's tau) of the correlation between allele divergence and number of bound peptides from intracellular proteins was subtracted from the coefficient of correlation with extracellular proteins (for MHC class I genes, expected to be more strongly associated with intracellular proteins) or vice versa (for MHC class II genes). Observed values (arrows) and 5% significance cut-off (red dotted lines) are also reported for the five HLA genes.



Supplementary figure S11. Mean amino-acid composition for each amino acid in two groups of pathogen proteins: extracellular (orange), intracellular (green) together with the statistical significance between each pair of groups. Bars sharing the same letter are not significantly different, while different letters indicate statistically significant differences between groups ($p < 0.05$).



Supplementary table S1. HLA-A, -B, -C, -DRB1 and -DQB1 alleles listed as '*common*' in the CWD catalogue (Mack et al. 2013) and used for binding prediction, sequence divergence and frequencies analysis. Two distinct phylogenetic groups of DRB1, denoted as group A and B, were also considered separately for the analysis.

HLA-A (n = 63)			HLA-B (n = 123)					HLA-C (n = 40)		HLA-DRB1 (n = 73)				HLA-DQB1 (n = 21)
										Group A (n = 46)		Group B (n = 27)		
*01:01	*11:02	*34:02	*07:02	*15:21	*35:14	*40:05	*51:02	*01:02	*12:04	*03:01	*13:01	*01:01	*16:04	*02:01
*01:02	*11:04	*34:05	*07:04	*15:24	*35:16	*40:06	*51:05	*02:02	*14:02	*03:02	*13:02	*01:02	*16:07	*02:02
*01:03	*23:01	*36:01	*07:05	*15:25	*35:17	*40:08	*51:06	*02:10	*14:03	*08:01	*13:03	*01:03		*02:03
*02:01	*24:02	*43:01	*07:06	*15:30	*35:19	*40:16	*51:07	*03:02	*15:02	*08:02	*13:04	*04:01		*03:01
*02:02	*24:03	*66:01	*07:10	*15:32	*35:20	*40:27	*51:08	*03:03	*15:04	*08:03	*13:05	*04:02		*03:02
*02:03	*24:07	*66:02	*08:01	*15:34	*35:21	*41:01	*51:09	*03:04	*15:05	*08:04	*13:07	*04:03		*03:03
*02:04	*24:10	*68:01	*13:01	*15:35	*35:23	*41:02	*52:01	*03:05	*15:16	*08:06	*13:10	*04:04		*03:04
*02:05	*24:17	*68:02	*13:02	*15:39	*35:41	*41:03	*53:01	*03:06	*15:09	*08:07	*13:12	*04:05		*03:05
*02:06	*24:20	*68:03	*14:01	*18:01	*35:43	*42:01	*54:01	*04:01	*16:01	*08:10	*13:14	*04:06		*03:09
*02:07	*24:25	*68:05	*14:02	*18:02	*37:01	*42:02	*55:01	*04:03	*16:02	*08:11	*13:20	*04:07		*03:19
*02:08	*24:33	*69:01	*14:03	*18:03	*38:01	*44:02	*55:02	*04:04	*16:04	*10:01	*14:01	*04:08		*04:01
*02:09	*25:01	*74:01	*15:01	*27:02	*38:02	*44:03	*56:01	*04:07	*17:01	*11:01	*14:02	*04:09		*04:02
*02:10	*26:01	*74:03	*15:02	*27:03	*39:01	*44:04	*57:01	*05:01	*17:03	*11:02	*14:03	*04:10		*05:01
*02:11	*26:03		*15:03	*27:04	*39:02	*44:05	*57:02	*06:02	*18:01	*11:03	*14:04	*04:11		*05:02
*02:14	*26:08		*15:07	*27:05	*39:03	*44:10	*57:03	*07:01	*18:02	*11:04	*14:05	*04:17		*05:03
*02:17	*29:01		*15:08	*27:06	*39:05	*44:27	*58:01	*07:02		*11:06	*14:06	*04:38		*06:01
*02:20	*29:02		*15:09	*27:07	*39:06	*45:01	*58:02	*07:04		*11:09	*14:07	*07:01		*06:02
*02:22	*30:01		*15:10	*27:08	*39:08	*46:01	*59:01	*07:06		*11:10	*14:17	*09:01		*06:03
*02:24	*30:02		*15:11	*35:01	*39:10	*47:01	*67:01	*07:18		*11:11	*14:18	*15:01		*06:04
*02:30	*30:04		*15:12	*35:02	*39:11	*48:01	*73:01	*08:01		*11:12	*14:24	*15:02		*06:08
*02:60	*31:01		*15:13	*35:03	*39:12	*48:03	*78:01	*08:02		*11:15	*14:54	*15:03		*06:09
*03:01	*32:01		*15:15	*35:04	*39:15	*49:01	*81:01	*08:03		*11:17		*15:04		
*03:02	*33:01		*15:16	*35:05	*39:24	*50:01	*82:01	*08:04		*11:19		*15:06		
*03:05	*33:03		*15:17	*35:08	*40:01	*50:02		*12:02		*12:01		*16:01		
*11:01	*34:01		*15:18	*35:12	*40:02	*51:01		*12:03		*12:02		*16:02		

Supplementary table S2. Classification table for pathogens categories and pathogen protein's accession numbers.

Classification	Pathogen	Protein name	Accession number
Extracellular (n = 57)	Clostridium tetani	tetanus toxin tetX (plasmid)	NP_783831
		tetanolysin O	YP_008774065
		hemolysin	NP_782184
		fibronectin-binding protein	NP_780878
		S-layer protein/N-acetylmuramoyl-L-alanine amidase	NP_781182
	Vibrio cholerae	cholera enterotoxin subunit B	NP_231099
		cholera enterotoxin subunit A	NP_231100
		accessory cholera enterotoxin	NP_231102
		colonization factor	NP_231104
	Entamoeba histolytica	calreticulin	XP_655241
		membrane-bound O-acyltransferase (MBOAT) family protein	EAL50306
		membrane transporter, putative	EAL50995
		adhesin 112 (EhADH112)	XP_652992
		surface antigen ariel1	AAC72364
		guanine nucleotide-binding protein subunit beta 2-like 1	XP_657050
		enolase	XP_649161
		glyceraldehyde-phosphate dehydrogenase, partial	AAA29100
		triosephosphate isomerase	XP_650725
		EF-hand calcium-binding domain containing protein	XP_648032
	Giardia lamblia	Seven transmembrane protein 1	XP_001704065
		High cysteine membrane protein Group 6	XP_001705852
		High cysteine membrane protein VSP-like	XP_001705828
		VSP, putative	XP_001703888
		VSP	XP_001705844
		VSP	XP_001705829
		VSP	XP_001705785
		VSP	P_001705784
		VSP	XP_001705783

		VSP	XP_001703933
		VSP	XP_001703925
		VSP	XP_001703844
		VSP	XP_001703931
	Trichinella spiralis	heat shock protein 70	CAA73574
		cysteine protease ATG4B	EFV52545
		14-3-3 protein	ACV51809
		serine protease inhibitor Kazal-type 4	EFV53657
	Trichomonas vaginalis	P270 surface immunogen-like, partial	XP_001292151
		surface antigen BspA-like	XP_001313891
		surface immunogen P270-related protein, partial	XP_001284871
		alpha-actinin	XP_001580136
		hypothetical protein	XP_001582296
		Ser/Thr protein phosphatase	XP_001299798
		Adenylate and Guanylate cyclase catalytic domain containing protein	XP_001584317
		Tetraspanin family protein	XP_001327241
		polymorphic outer membrane protein	XP_001325298
		polymorphic outer membrane protein	XP_001303631
		Clan S-, family S54, Rhomboid-like serine peptidase	XP_001316801
		Tetraspanin family protein	XP_001326883
		hypothetical protein	XP_001301868
	Schistosoma mansoni	putative paramyosin (Allergen Blo t 11)	CCD74732
		putative paramyosin (Allergen Blo t 11)	CCD75328
		glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	CCD75626
		glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	CCD75625
		glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	CCD75627
		glutathione S-transferase 26 kDa	CCD80234
		putative fatty acid binding protein	CCD77656
		putative fatty acid binding protein	CCD77655
Intracellular (n = 100)	Human immunodeficiency virus 1	Gag-Pol	NP_057849
		Pr55(Gag)	NP_057850

		Vif	NP_057851
		Vpr	NP_057852
		Tat	NP_057853
		Rev	NP_057854
		Vpu	NP_057855
		Envelope surface glycoprotein gp160, precursor	NP_057856
		Asp	YP_009028572
		Nef	NP_057857
	Hepatitis B virus	Polymerase	NP_647604
		large envelope protein	YP_355333
		middle envelope protein	YP_355334
		small envelope protein	NP_647605
		X protein	NP_647606
		pre-capsid protein	YP_355335
		capsid protein	NP_647607
	Measles virus	nucleocapsid protein	NP_056918
		phosphoprotein	NP_056919
		V protein	YP_003873249
		C protein	NP_056920
		matrix protein	NP_056921
		fusion protein	NP_056922
		hemagglutinin protein	NP_056923
		large polymerase protein	NP_056924
	Mumps virus	nucleocapsid protein	NP_054707
		phoshoprotein	NP_054708
		V protein	NP_054709
		membrane protein	NP_054710
		fusion protein	NP_054711
		small hydrophobic protein	NP_054712
		hemagglutinin-neuraminidase	NP_054713
		large protein	NP_054714

	Rabies virus	nucleoprotein N	NP_056793
		phosphoprotein M1	NP_056794
		M2 protein	NP_056795
		transmembrane glycoprotein G	NP_056796
		L protein	NP_056797
	Rubella virus	non-structural polyprotein	NP_062883
		structural polyprotein	NP_062884
	Hepatitis C virus	polyprotein	NP_671491
		protein F	NP_803170
	Human herpesvirus 4	early antigen protein D	YP_401657
		envelope glycoprotein gp350	YP_401667
		glycoprotein L precursor	YP_401678
		glycoprotein M	YP_401685
		putative membrane antigen gp85	YP_401694
		probable membrane glycoprotein	YP_401706
		glycoprotein gp85 precursor	YP_401700
		glycoprotein gp110 precursor	YP_401713
		putative nmembrane antigen p140	YP_401633
		membrane protein	YP_401711
		putative membrane protein BLRF1	YP_401665
		large tegument protein	YP_401652
		EBNA-1 protein	YP_401677
		EBNA-2 nuclear protein	YP_401644
		EBNA-LP protein	YP_401636
		EBNA3A nuclear protein	YP_401669
		EBNA3C latent protein	YP_401671
		EBNA-3B nuclear protein	YP_401670
		terminal protein LMP2A	YP_401631
		terminal protein LMP2B	YP_401632
		BCRF1 protein precursor	YP_401634
		hypothetical protein BCRF2, partial	YP_401635

	BMRF2 protein	YP_401658
	BMLF1 protein	YP_401660
	putative BLLF2 protein	YP_401668
	putative BZLF2 protein	YP_401672
	putative BARF0 protein, partial	YP_401716
Variola virus	A-type inclusion body protein	NP_042174
	putative virulence factor	NP_042060
	36 kDa major membrane protein	NP_042073
	palmytilated EEV membrane glycoprotein	NP_042081
	IMV membrane protein	NP_042100
	putative viral membrane protein	NP_042129
	IMV membrane protein	NP_042168
	IEV and EEV membrane glycoprotein	NP_042185
	EEV membrane glycoprotein	NP_042219
Influenza A virus	nucleocapsid protein	AJI80522
	hemagglutinin	AJI80518
	neuraminidase	AJI80521
Treponema pallidum	antigen	AAA75016
	treponemal membrane protein A	AAA27481
	endoflagellar sheath protein	AAA27477
	17 kDa lipoprotein	AAA27472
	Tpp15	AAC45732
Mycobacterium leprae	FAP family protein	NP_302372
	cytotoxin/hemolysin	NP_301968
	p60-family protein	NP_301879
	co-chaperonin GroES	NP_301372
	putative secreted protein	CAR70980
Toxoplasma gondii	hypothetical protein TGME49_279540	EPT30382
	dense granule protein GRA7	EPT30138
	aspartyl protease ASP1	EPT30276
	rhoptry protein ROP18	EPT29989

		MIC2-associated protein M2AP	EPT26499
		dense granule protein GRA2	EPT27242
		dense granule protein GRA6	EPT26403
		SAG-related sequence SRS39	EPT30357
		SAG-related sequence SRS35A	EPT29845
Intra–extracellular (n = 75)	Mycobacterium tuberculosis	ESAT-6-like protein EsxB	NP_218391
		ribonucleoside-diphosphate reductase subunit beta NrdF1	YP_177853
		PPE family protein PPE42	YP_177893
		NAD(P)H nitroreductase	NP_216548
		ESX-1 secretion-associated protein EspI	NP_218393
		PPE family protein PPE68	YP_178022
		ESAT-6 like protein EsxJ	NP_215554
		diacylglycerol acyltransferase/mycolyltransferase Ag85A	NP_218321
		heparin binding hemagglutinin HbhA	NP_214989
		permease	NP_216495
	Streptococcus pneumoniae	surface protein pspA	NP_357715
		choline binding protein A	P_359586
		N-acetylmuramoyl-L-alanine amidase	NP_359346
		endo-beta-N-acetylglucosaminidase	NP_358461
		1,4-beta-N-acetylmuramidase	NP_359024
		Para-aminobenzoate synthetase	NP_358176
		neuraminidase A	NP_359129
		manganese ABC transporter substrate-binding protein	NP_359087
		cell wall-associated serine proteinase PrtA	NP_358155
		zinc metalloprotease	NP_358175
	Yersinia pestis	F1 capsule antigen (plasmid)	NP_395430
		putative F1 capsule anchoring protein	NP_395429
		putative F1 operon positive regulatory protein (plasmid)	NP_395427
		secreted effector protein (plasmid)	NP_395165
		low calcium response protein G (plasmid)	NP_395166
		putative outer membrane virulence protein (plasmid)	NP_395143

	Bordetella pertussis	pertussis toxin subunit 1	NP_882282
		pertussis toxin subunit 2	NP_882283
		pertussis toxin subunit 3	NP_882286
		pertussis toxin subunit 4	NP_882284
		pertussis toxin subunit 5	NP_882285
		serotype 2 fimbrial subunit	NP_879898
		serotype 3 fimbrial subunit	NP_880302
		filamentous hemagglutinin/adhesin	NP_880571
		pertactin autotransporter	NP_879839
		dermonecrotic toxin	NP_881965
	Corynebacterium diphtheriae	diphtheria toxin	NP_938615
	Salmonella enterica	type III secretion protein	NP_456106
		type III secretion protein	NP_456107
		type III secretion protein	NP_456108
		type III secretion protein	NP_456109
		type III secretion protein	NP_456110
		type III secretion protein	NP_456114
		outer membrane secretory protein	NP_456134
		pathogenicity island 2 secreted effector protein	NP_456135
		two-component sensor kinase	NP_456136
		Vi polysaccharide exporter protein	NP_458730
		Vi polysaccharide exporter inner-membrane protein	NP_458731
		Vi polysaccharide exporter protein	NP_458734
		Vi polysaccharide exporter ATP-binding protein	NP_458732
		Vi polysaccharide exporter inner-membrane protein	NP_458733
		Vi polysaccharide biosynthesis protein	NP_458738
		Vi polysaccharide biosynthesis epimerase	NP_458739
		Vi polysaccharide biosynthesis protein	NP_458741
		Vi polysaccharide biosynthesis protein TviE	NP_458735
	Plasmodium falciparum	rhoptry-associated protein 1, RAP1	XP_001348275
		conserved Plasmodium protein, unknown function	XP_001348247

	reticulocyte binding protein 2 homolog A	XP_001350083
	reticulocyte binding protein 2, homolog b	XP_002809051
	plasmodium falciparum gamete antigen 27/25	XP_001349749
	Thrombospondin-related anonymous protein, TRAP	XP_001350088
	erythrocyte binding antigen-140	XP_001349859
	CelTOS, putative	XP_001350569
	erythrocyte membrane protein 1, PfEMP1	XP_001350410
	MAEBL, putative (merozoite adhesive erythrocytic binding protein)	XP_001348153
	apical membrane antigen 1, AMA1	XP_001348015
	circumsporozoite-related antigen	XP_001347895
	merozoite surface protein	XP_001347636
	merozoite surface protein 6	XP_001347630
	merozoite surface protein 3	XP_001347629
	cytoadherence linked asexual protein 9(CLAG9)	XP_001352222
	merozoite surface protein 1 precursor	XP_001352170
	heat shock 70 kDa protein	XP_001349336
	reticulocyte-binding protein homologue 1	XP_002808637
	merozoite surface protein 2 precursor	XP_001349578

Supplementary table S3. Correlation values between pairwise parameters of allele divergence.

	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1	
	tau	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**
p-distance ~ jtt	0.977	< 0.001	0.974	< 0.001	0.952	< 0.001	0.962	< 0.001	0.975	< 0.001
p-distance ~ dayhoff	0.94	< 0.001	0.933	< 0.001	0.931	< 0.001	0.922	< 0.001	0.956	< 0.001
p-distance ~ grantham	0.859	< 0.001	0.855	< 0.001	0.744	< 0.001	0.845	< 0.001	0.929	< 0.001
p-distance ~ sandberg	0.886	< 0.001	0.847	< 0.001	0.823	< 0.001	0.879	< 0.001	0.928	< 0.001
jtt ~dayhoff	0.931	< 0.001	0.925	< 0.001	0.914	< 0.001	0.923	< 0.001	0.957	< 0.001
jtt ~ grantham	0.843	< 0.001	0.839	< 0.001	0.702	< 0.001	0.824	< 0.001	0.914	< 0.001
jtt ~ sandberg	0.875	< 0.001	0.832	< 0.001	0.785	< 0.001	0.855	< 0.001	0.914	< 0.001
dayhoff ~ grantham	0.875	< 0.001	0.85	< 0.001	0.724	< 0.001	0.844	< 0.001	0.909	< 0.001
dayhoff ~ sandberg	0.896	< 0.001	0.828	< 0.001	0.79	< 0.001	0.876	< 0.001	0.915	< 0.001
grantham ~ sandberg	0.923	< 0.001	0.884	< 0.001	0.808	< 0.001	0.904	< 0.001	0.947	< 0.001

tau: Kendall's tau coefficient; p_{adj}: p-value after Bonferroni-correction across multiple alleles tested at each locus and number of loci

Supplementary table S4. Correlation values between combined number of bound peptide and genetic distance between all possible alleles pairs across the five key classical MHC genes. Genetic distances determined using five different pairwise parameters of allele divergence: p-distance, DayHoff and JTT, Grantham and Sandberg. Binding prediction was performed on the complete dataset of pathogen proteins (n=232) as well as considering proteins separately within three groups of pathogens: extracellular (n = 57), intracellular (n = 100) and intra-extracellular (n = 75).

		HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1	
Pathogen groups	**Parameters**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**
Total	P-distance	0.420	< 0.001	0.375	< 0.001	0.412	< 0.001	0.174	< 0.001	0.223	< 0.001
	Dayhoff	0.387	< 0.001	0.364	< 0.001	0.415	< 0.001	0.174	< 0.001	0.226	< 0.001
	Jtt	0.410	< 0.001	0.366	< 0.001	0.377	< 0.001	0.179	< 0.001	0.215	< 0.001
	Grantham	0.361	< 0.001	0.397	< 0.001	0.507	< 0.001	0.157	< 0.001	0.210	0.001
	Sandberg	0.380	< 0.001	0.361	< 0.001	0.438	< 0.001	0.165	< 0.001	0.206	0.001
Extracellular	P-distance	0.394	< 0.001	0.148	< 0.001	0.333	< 0.001	0.136	< 0.001	0.316	< 0.001
	Dayhoff	0.365	< 0.001	0.127	< 0.001	0.325	< 0.001	0.149	< 0.001	0.321	< 0.001
	Jtt	0.383	< 0.001	0.143	< 0.001	0.300	< 0.001	0.145	< 0.001	0.310	< 0.001
	Grantham	0.345	< 0.001	0.192	< 0.001	0.392	< 0.001	0.130	< 0.001	0.303	< 0.001
	Sandberg	0.355	< 0.001	0.194	< 0.001	0.362	< 0.001	0.135	< 0.001	0.299	< 0.001
Extra-Intra	P-distance	0.413	< 0.001	0.267	< 0.001	0.384	< 0.001	0.195	< 0.001	0.223	0.010
	Dayhoff	0.375	< 0.001	0.246	< 0.001	0.378	< 0.001	0.196	< 0.001	0.229	0.016
	Jtt	0.402	< 0.001	0.255	< 0.001	0.352	< 0.001	0.201	< 0.001	0.217	0.007
	Grantham	0.351	< 0.001	0.289	< 0.001	0.475	< 0.001	0.166	< 0.001	0.210	0.024
	Sandberg	0.370	< 0.001	0.276	< 0.001	0.415	< 0.001	0.175	< 0.001	0.206	0.029
Intracellular	P-distance	0.325	< 0.001	0.387	< 0.001	0.439	< 0.001	0.149	< 0.001	0.187	< 0.001
	Dayhoff	0.314	< 0.001	0.397	< 0.001	0.450	< 0.001	0.140	< 0.001	0.186	< 0.001
	Jtt	0.321	< 0.001	0.383	< 0.001	0.405	< 0.001	0.150	< 0.001	0.177	< 0.001
	Grantham	0.293	< 0.001	0.377	< 0.001	0.544	< 0.001	0.137	< 0.001	0.172	0.001
	Sandberg	0.314	< 0.001	0.326	< 0.001	0.454	< 0.001	0.143	< 0.001	0.170	0.001

tau: Kendall's tau coefficient; p_{adj}: p-value after Bonferroni-correction across multiple alleles tested at each locus and number of loci

Supplementary table S5. Correlation values between combined number of bound peptide and Grantham genetic distance between all possible alleles pairs across the two class II loci (HLA-DRB1 and HLA-DQB1). Binding prediction analysis of all possible 15mer pathogen-derived peptides were performed considering the complete dataset of pathogen proteins (n=232) as well as considering proteins separately within three groups of pathogens: extracellular (n = 57), intracellular (n = 100) and intra-extracellular (n = 75).

	HLA-DRB1		**HLA-DQB1**	
Pathogen groups	**tau**	**p_{adj}**	**tau**	**p_{adj}**
Total	0.355	< 0.001	0.135	0.401
Extracellular	0.271	< 0.001	0.367	< 0.001
Extra-Intra	0.332	< 0.001	0.086	6.994
Intracellular	0.34	< 0.001	0.112	1.720

tau: Kendall's tau coefficient; p_{adj}: p-value after Bonferroni-correction across multiple alleles tested at each locus and number of loci

Supplementary table S6. Correlation values between combined number of bound peptide and Grantham genetic distance between all possible alleles pairs across the five key classical MHC genes. Binding prediction analysis were performed using the established binding threshold (%rank of 0.5) which includes only strong binders, considering the complete dataset of pathogen proteins (n=232) as well as considering proteins separately within three groups of pathogens: extracellular (n = 57), intracellular (n = 100) and intra-extracellular (n = 75).

	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1	
Pathogen groups	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**
Total	0.132	< 0.001	0.225	< 0.001	0.400	< 0.001	0.086	< 0.001	0.182	0.011
Extracellular	0.154	< 0.001	0.088	< 0.001	0.267	< 0.001	0.082	< 0.001	0.241	< 0.001
Extra-Intra	0.160	< 0.001	0.130	< 0.001	0.397	< 0.001	0.076	< 0.001	0.209	0.001
Intracellular	0.091	< 0.001	0.217	< 0.001	0.395	< 0.001	0.075	< 0.001	0.146	0.206

tau: Kendall's tau coefficient; p_{adj}: p-value after Bonferroni-correction across multiple alleles tested at each locus and number of loci

Supplementary table S7. Correlations values between combined number of bound peptide and Grantham genetic distance between all possible allele pairs at each of the five key classical MHC genes, reported for four different sets of artificial proteins.

		Original proteins		AAC within proteins		AAC within pathogen groups		AAC across all pathogens		AAC UniProtKB/Swiss-Prot	
Loci	**Pathogen groups**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**	**tau**	**p_{adj}**
HLA-A	Extracellular	0.345	< 0.001	0.317	< 0.001	0.197	< 0.001	0.153	< 0.001	0.096	< 0.001
	Extra-Intra	0.351	< 0.001	0.225	< 0.001	0.141	< 0.001	0.125	< 0.001	0.106	< 0.001
	Intracellular	0.293	< 0.001	0.222	< 0.001	0.139	< 0.001	0.163	< 0.001	0.126	< 0.001
HLA-B	Extracellular	0.192	< 0.001	0.237	< 0.001	0.196	< 0.001	0.314	< 0.001	0.309	< 0.001
	Extra-Intra	0.289	< 0.001	0.252	< 0.001	0.235	< 0.001	0.330	< 0.001	0.343	< 0.001
	Intracellular	0.377	< 0.001	0.375	< 0.001	0.366	< 0.001	0.333	< 0.001	0.356	< 0.001
HLA-C	Extracellular	0.392	< 0.001	0.361	< 0.001	0.317	< 0.001	0.495	< 0.001	0.450	< 0.001
	Extra-Intra	0.475	< 0.001	0.465	< 0.001	0.436	< 0.001	0.495	< 0.001	0.499	< 0.001
	Intracellular	0.544	< 0.001	0.531	< 0.001	0.543	< 0.001	0.491	< 0.001	0.493	< 0.001
HLA-DRB1	Extracellular	0.130	< 0.001	0.151	< 0.001	0.136	< 0.001	0.177	< 0.001	0.131	< 0.001
	Extra-Intra	0.166	< 0.001	0.144	< 0.001	0.129	< 0.001	0.169	< 0.001	0.198	< 0.001
	Intracellular	0.137	< 0.001	0.249	< 0.001	0.186	< 0.001	0.168	< 0.001	0.165	< 0.001
HLA-DQB1	Extracellular	0.303	< 0.001	0.407	< 0.001	0.414	< 0.001	0.367	< 0.001	0.420	< 0.001
	Extra-Intra	0.210	0.024	0.439	< 0.001	0.435	< 0.001	0.418	< 0.001	0.419	< 0.001
	Intracellular	0.172	0.001	0.349	< 0.001	0.383	< 0.001	0.399	< 0.001	0.438	< 0.001

tau: Kendall's tau coefficient; p_{adj}: p-value after Bonferroni-correction across multiple alleles tested at each locus and number of loci

Peptide binding prediction was performed for different sets of artificial proteins in order to test for a potential bias by certain HLA genes towards specific pathogen groups and estimate the effect of group-specific amino acid composition (AAC). Artificial proteins were created in four different ways: by randomly shuffling the amino acids within each pathogen protein (AAC within proteins), by maintaining the amino acid frequencies as they occur within each group of pathogen proteins (AAC within pathogen groups), by considering amino acid frequencies as they occur in the whole dataset of pathogenic proteins (AAC across all pathogens), or by using the amino acid frequencies computed from UniProtKB/Swiss-Prot data bank (AAC UniProtKB/Swiss-Prot).

Supplementary table S8. Comparison of average amino acid composition between the two groups of pathogen proteins (extracellular and intracellular) performed using one-way analysis of variance. Mean amino-acid composition and standard deviation within the two groups of pathogen proteins are reported for each amino acid, together with the statistical significance between each pair of groups.

	Extracellular	**Intracellular**	**p_{adj}**
A (Ala)	7.14 (2.89)	7.86 (3.35)	3.540
R (Arg)	3.04 (1.71)	6.17 (3.01)	< 0.001
N (Asn)	6 (2.78)	3.58 (2.07)	< 0.001
D (Asp)	5.63 (1.69)	4.44 (2)	0.004
C (Cys)	3.79 (4.31)	2.23 (1.72)	0.032
Q (Gln)	3.02 (1.8)	3.92 (2.19)	0.182
E (Glu)	5.67 (2.86)	5.19 (2.53)	5.600
G (Gly)	7.18 (3.56)	7.32 (4.21)	16.600
H (His)	1.39 (1.08)	1.99 (1.33)	0.080
I (Ile)	6.86 (3.22)	5.09 (2.91)	0.011
L (Leu)	7.3 (3.01)	9.74 (3.65)	0.001
K (Lys)	7.14 (2.46)	4.29 (2.59)	< 0.001
M (Met)	2.25 (1.62)	2.22 (1.19)	18.200
F (Phe)	4.09 (1.93)	3.5 (1.87)	1.260
P (Pro)	3.6 (1.7)	6.85 (4.55)	< 0.001
S (Ser)	7.51 (3.01)	7.98 (2.28)	5.400
T (Thr)	6.56 (2.57)	6.2 (2.88)	8.600
W (Trp)	1.04 (1.03)	1.55 (1.24)	0.178
Y (Tyr)	3.49 (1.85)	2.76 (1.46)	0.140
V (Val)	7.12 (2.83)	6.84 (2.33)	10.000

p_{adj}: p-value after Bonferroni-correction across the number of amino acids tested

Supplementary table S9. Correlations values between the average Grantham pairwise divergence to the most common alleles and the allele frequency in different European populations for four classical HLA loci. USA NMDP: HLA allele frequencies information provided from the National Marrow Donor Program.

	HLA-A			HLA-B			HLA-C			HLA-DRB1		
population	**tau**	**P**	**p_{adj}**	**tau**	**P**	**p_{adj}**	**tau**	**P**	**p_{adj}**	**tau**	**P**	**p_{adj}**
USA NMDP European Caucasian	0.110	0.255	0.766	0.241	< 0.001	0.001	0.235	0.119	0.356	0.194	0.024	0.072
German	0.150	0.120	0.356	0.173	0.012	0.036	0.280	0.018	0.054	0.254	0.004	0.012
Poland	0.061	0.573	1.722	0.103	0.189	0.564	0.264	0.038	0.113	0.201	0.035	0.104

tau: Kendall's tau coefficient; P: p-value from Kendall correlation; p_{adj}: p-value after Bonferroni-correction across multiple tested populations