SUPPLEMENTARY INFORMATION

Supplementary Notes 1: accuracy of prediction algorithms for peptide binding affinities to HLA and Mamu alleles

For each HLA and Mamu allele we have analyzed the accuracy of four predictive algorithms available from the Immune Epitope Database (IEDB): ANN and SMM (versions 2009-09-01 and 2007-12-27). Accuracy was tested against experimental data (downloaded from the IEDB) of measured peptide binding affinities (IC_{50}) to HLA and Mamu molecules. Only those HLA and Mamu alleles were kept in the analysis for which there was enough experimental data (at least 50 binders and 50 non-binders).

First we tested how good the predictive algorithms are in classifying peptides into binders (IC_{50} < 500 nM) and non-binders (IC_{50}≥500 nM). We counted the number of true positives *TP* (correctly predicted binders), true negatives *TN* (correctly predicted non-binders), false positives *FP* (incorrectly predicted binders) and false negatives *FN* (incorrectly predicted non-binders). The accuracy of the algorithm is defined as

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}. \qquad \text{(S1)}$$

Most algorithms for which there were sufficient experimental data were very accurate (more than 80%, Table S1). Commonly used measure of accuracy is also Matthews correlation coefficient:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot M \cdot P' \cdot M'}}, \qquad \text{(S2)}$$

where *P (M)* and *P' (M')* are numbers of experimental binders (non-binders) and predicted binders (non-binders) respectively. The closer the *MCC* is to the value 1, the higher the accuracy of the prediction algorithm. Most algorithms had a *MCC* value in the range of 0.6 to 0.9 (Table S1).

Second we tested how good the predictive algorithms are at determining the actual value of the binding affinity, IC_{50}. Because binding affinities span a huge range of values, a commonly used difference of logarithms

$$f(i) = \ln(\text{predicted } IC_{50}) - \ln(\text{measured } IC_{50}), \qquad \text{(S3)}$$

was taken as a measure of the accuracy of predicted binding affinity of the *i*-th peptide. When experiments reported that binding affinity has a value greater than some value, LIC_{50}, we defined

$$f(i) = \begin{cases} 0; & \text{predicted } IC_{50} \geq LIC_{50} \\ \ln(\text{predicted } IC_{50}) - \ln(LIC_{50}); & \text{predicted } IC_{50} < LIC_{50} \end{cases}. \quad \text{(S4)}$$

Similarly, when experiments reported that binding affinity has value lower than some value, HIC_{50}, the accuracy was defined as

$$f(i) = \begin{cases} 0; & \text{predicted } IC_{50} \leq HIC_{50} \\ \ln(\text{predicted } IC_{50}) - \ln(HIC_{50}); & \text{predicted } IC_{50} < HIC_{50} \end{cases}. \quad \text{(S5)}$$

Overall accuracy of the predictive algorithm was determined from average bias $\Delta\ln(y)$ and average root mean square error σ:

$$\Delta\ln(y) = \frac{1}{N}\sum_{i=1}^{N} f(i), \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left[f(i)\right]^2. \quad \text{(S6)}$$

Table S1 reports accuracies of all 4 predictive algorithms for each HLA-B allele and Mamu allele for which there was enough experimental data available. HLA-A alleles were not studied, because they are not associated with control of HIV[23]. In assessing the accuracy of the algorithms using Eqs. S1-S6, we did not include experimental data for which even a bound (less than or greater than) for IC_{50} was not reported.

If average bias $\Delta\ln(y) > 0$ ($\Delta\ln(y) < 0$), then the predictive algorithm on average overestimates (underestimates) the value of binding affinity. Average bias of predictive algorithms can affect values of predicted fraction of peptides that can bind to a certain allele. In Table S1, alleles for which there are no accurate predictive algorithms (all 4 predictive algorithms have large normalized bias $|\Delta\ln(y)/\sigma| > 0.06$) are marked with white. The most accurate algorithm (bold font in Table S1) for those alleles is taken to be the one with the least value of normalized bias ($|\Delta\ln(y)/\sigma|$). In Table S1, alleles for which there is at least one accurate predictive algorithm (normalized bias $|\Delta\ln(y)/\sigma| < 0.06$) are marked with yellow. If there is more than one accurate predictive algorithm for a given allele, the most accurate predictive algorithm (bold font in Table S1) was selected to be the one with the least value of average root mean square error , σ, among the accurate predictive algorithms (for all of which $|\Delta\ln(y)/\sigma|$ was very small).

Different choices of threshold for normalized bias that separate accurate and inaccurate predictive algorithms lead to only small changes in Table S1. For example, if we choose the threshold to be 0.08, we get one more allele (HLA-B*1517) with at least one accurate predictive algorithm and the most accurate algorithm would change for two alleles (HLA-B*1517 and HLA-B*4403).

For each HLA and Mamu allele we used the most accurate predictive algorithm to predict the fraction of peptides derived from human and macaque proteome that can bind to given allele. There were ~10^7 (~10^6) unique peptide sequences in human (macaque) proteome. We used only HLA-B alleles (marked with yellow in Table S1) for which there is at least one accurate predictive algorithm available to determine the typical binding fraction for HLA-B alleles (median – 0.013 and average – 0.015).

Supplementary Table S1:

Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles

allele	BF	predictive algorithm	N	ACC	MCC	$\Delta \ln(y)$	σ	$\lvert\Delta \ln(y)/\sigma\rvert$
HLA-B*0702	0.016	ann (2009-09-01)	2301	94.1	0.840	0.108	1.065	0.101
	0.018	**ann (2007-12-27)**		**93.9**	**0.841**	**-0.061**	**1.074**	**0.057**
	0.017	smm (2009-09-01)		92.8	0.807	0.150	1.419	0.105
	0.024	smm (2007-12-27)		92.3	0.795	0.079	1.409	0.056
HLA-B*0801	0.011	ann (2009-09-01)	1560	91.2	0.759	0.246	1.553	0.159
	0.023	ann (2007-12-27)		89.5	0.744	-0.260	1.603	0.162
	0.012	smm (2009-09-01)		89.0	0.701	0.254	1.733	0.147
	0.038	**smm (2007-12-27)**		**88.9**	**0.724**	**-0.191**	**1.700**	**0.112**
HLA-B*1501	0.028	ann (2009-09-01)	2342	87.3	0.713	0.243	1.519	0.160
	0.035	**ann (2007-12-27)**		**87.8**	**0.727**	**-0.061**	**1.470**	**0.041**
	0.023	smm (2009-09-01)		85.4	0.668	0.248	1.640	0.151
	0.036	smm (2007-12-27)		86.4	0.694	0.005	1.564	0.003
HLA-B*1503	0.068	ann (2009-09-01)	390	88.2	0.586	0.699	1.934	0.361
	0.077	ann (2007-12-27)		89.2	0.598	0.391	1.823	0.214
	0.313	smm (2009-09-01)		88.2	0.541	0.301	1.789	0.168
	0.451	**smm (2007-12-27)**		**89.5**	**0.580**	**0.266**	**1.898**	**0.140**
HLA-B*1517	0.050	ann (2009-09-01)	678	95.4	0.904	0.234	1.244	0.188
	0.059	ann (2007-12-27)		95.0	0.895	-0.104	1.323	0.078
	0.079	smm (2009-09-01)		93.5	0.863	0.125	1.437	0.087
	0.141	**smm (2007-12-27)**		**94.1**	**0.877**	**-0.105**	**1.450**	**0.072**
HLA-B*1801	**0.009**	**ann (2009-09-01)**	1161	**94.7**	**0.730**	**-0.034**	**1.081**	**0.031**
	0.008	ann (2007-12-27)		94.9	0.760	-0.205	1.110	0.185
	0.010	smm (2009-09-01)		93.5	0.667	-0.004	1.329	0.003
	0.012	smm (2007-12-27)		94.0	0.707	-0.120	1.355	0.089
HLA-B*2705	**0.014**	**ann (2009-09-01)**	1701	**95.3**	**0.792**	**-0.024**	**0.775**	**0.031**
	0.017	ann (2007-12-27)		94.9	0.797	-0.106	0.843	0.125
	0.016	smm (2009-09-01)		93.8	0.728	-0.028	0.972	0.029
	0.017	smm (2007-12-27)		93.5	0.714	-0.013	0.988	0.013

BF – fraction of 9-mer peptides (derived from human and macaque proteome) that are predicted to bind to HLA-B and Mamu alleles
N – number of available experimental measurements used to test prediction algorithms
ACC – % accuracy of classifying peptides into binders and non-binders (Eq. S1)
MCC – Matthews correlation coefficient (Eq. S2)
$\Delta \ln(y)$, σ – average bias and error of predicted binding affinity value IC_{50} (Eq. S6)

Detailed description of each quantity is available in Supplementary Notes 1. HLA-B and Mamu alleles for which prediction algorithms are sufficiently accurate (Supplementary Notes 1) are highlighted with yellow color. The most accurate predictive algorithm is marked with bold font.

Supplementary Table S1: continued

Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles

allele	BF	predictive algorithm	N	ACC	MCC	$\Delta\ln(y)$	σ	$\lvert\Delta\ln(y)/\sigma\rvert$
HLA-B*3501	0.024	ann (2009-09-01)	652	79.8	0.611	0.455	1.893	0.241
	0.027	**ann (2007-12-27)**		**80.8**	**0.622**	**0.203**	**2.041**	**0.100**
	0.034	smm (2009-09-01)		77.5	0.558	0.476	2.129	0.224
	0.042	smm (2007-12-27)		77.6	0.566	0.482	2.104	0.229
HLA-B*3901	**0.016**	**ann (2009-09-01)**	478	**94.4**	**83.2**	**-0.088**	**0.861**	**0.103**
		ann (2007-12-27)						
	0.022	smm (2009-09-01)		92.7	78.3	-0.183	1.083	0.170
		smm (2007-12-27)						
HLA-B*4001	0.011	ann (2009-09-01)	1832	96.0	0.845	0.082	0.945	0.087
	0.010	**ann (2007-12-27)**		**95.2**	**0.814**	**-0.011**	**1.181**	**0.009**
	0.015	smm (2009-09-01)		94.2	0.783	0.078	1.320	0.059
	0.010	smm (2007-12-27)		93.2	0.728	0.222	1.455	0.153
HLA-B*4002	0.013	ann (2009-09-01)	256	91.8	0.837	0.335	1.217	0.275
	0.019	ann (2007-12-27)		81.6	0.640	-0.621	1.850	0.336
	0.019	**smm (2009-09-01)**		**84.4**	**0.686**	**0.061**	**1.637**	**0.037**
	0.028	smm (2007-12-27)		82.8	0.657	-0.219	1.763	0.124
HLA-B*4402	0.003	ann (2009-09-01)	1052	95.8	0.640	-0.124	0.937	0.133
	0.004	ann (2007-12-27)		96.6	0.742	-0.216	0.987	0.219
	0.001	**smm (2009-09-01)**		**94.5**	**0.456**	**-0.105**	**1.063**	**0.099**
		smm (2007-12-27)						
HLA-B*4403	0.006	ann (2009-09-01)	260	87.3	0.662	-0.103	1.382	0.075
	0.007	ann (2007-12-27)		86.5	0.659	-0.676	1.648	0.410
	0.006	**smm (2009-09-01)**		**86.5**	**0.640**	**-0.047**	**1.597**	**0.029**
	0.011	smm (2007-12-27)		81.9	0.535	-0.340	1.789	0.190
HLA-B*4501	0.013	ann (2009-09-01)	249	95.2	0.885	-0.077	0.944	0.081
	0.011	ann (2007-12-27)		87.1	0.700	-0.366	1.526	0.240
	0.012	**smm (2009-09-01)**		**89.2**	**0.736**	**0.022**	**1.534**	**0.015**
	0.025	smm (2007-12-27)		85.9	0.670	0.057	1.810	0.032

BF – fraction of 9-mer peptides (derived from human and macaque proteome) that are predicted to bind to HLA-B and Mamu alleles
N – number of available experimental measurements used to test prediction algorithms
ACC – % accuracy of classifying peptides into binders and non-binders (Eq. S1)
MCC – Matthews correlation coefficient (Eq. S2)
$\Delta\ln(y)$, σ – average bias and error of predicted binding affinity value IC_{50} (Eq. S6)

Detailed description of each quantity is available in Supplementary Notes 1. HLA-B and Mamu alleles for which prediction algorithms are sufficiently accurate (Supplementary Notes 1) are highlighted with yellow color. The most accurate predictive algorithm is marked with bold font.

Supplementary Table S1: continued

Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles

allele	BF	predictive algorithm	N	ACC	MCC	$\Delta\ln(y)$	σ	$\lvert\Delta\ln(y)/\sigma\rvert$
HLA-B*5101	0.002	ann (2009-09-01)	849	89.4	0.609	0.160	1.461	0.109
	0.004	ann (2007-12-27)		89.3	0.669	-0.353	1.601	0.221
	0.001	**smm (2009-09-01)**		**87.8**	**0.531**	**0.158**	**1.675**	**0.094**
	0.007	smm (2007-12-27)		87.8	0.594	-0.185	1.651	0.112
HLA-B*5301	**0.008**	**ann (2009-09-01)**	399	**92.5**	**0.848**	**0.052**	**1.636**	**0.032**
	0.006	ann (2007-12-27)		89.0	0.780	-0.228	1.745	0.130
	0.017	smm (2009-09-01)		86.7	0.732	0.210	1.993	0.105
	0.025	smm (2007-12-27)		86.0	0.721	0.165	1.908	0.086
HLA-B*5401	0.010	ann (2009-09-01)	404	91.8	0.818	0.313	1.414	0.222
	0.009	ann (2007-12-27)		87.1	0.729	-0.370	1.742	0.212
	0.024	smm (2009-09-01)		89.1	0.754	0.174	1.775	0.098
	0.027	**smm (2007-12-27)**		**87.4**	**0.723**	**0.015**	**1.786**	**0.008**
HLA-B*5701	**0.007**	**ann (2009-09-01)**	1162	**96.6**	**0.837**	**0.000**	**0.640**	**0.001**
	0.008	ann (2007-12-27)		95.2	0.777	-0.230	0.901	0.255
	0.005	smm (2009-09-01)		94.1	0.696	-0.045	0.899	0.050
	0.006	smm (2007-12-27)		94.5	0.723	-0.087	0.871	0.100
HLA-B*5801	**0.016**	**ann (2009-09-01)**	1947	**95.3**	**0.838**	**0.052**	**0.956**	**0.054**
	0.012	ann (2007-12-27)		95.0	0.830	0.016	1.010	0.016
	0.017	smm (2009-09-01)		94.0	0.792	-0.039	1.194	0.033
	0.014	smm (2007-12-27)		93.9	0.788	0.093	1.205	0.077
Mamu-A*01	0.020	ann (2009-09-01)	692	87.4	0.749	0.150	1.597	0.094
	0.021	ann (2007-12-27)		85.8	0.718	-0.201	1.768	0.114
	0.028	smm (2009-09-01)		85.5	0.712	-0.001	1.909	0.001
	0.028	**smm (2007-12-27)**		**85.0**	**0.701**	**0.027**	**1.892**	**0.014**
Mamu-A*02	0.046	ann (2009-09-01)	249	83.5	0.677	0.421	2.135	0.197
	0.031	ann (2007-12-27)		82.3	0.646	0.060	2.322	0.026
	0.064	**smm (2009-09-01)**		**82.3**	**0.645**	**0.109**	**2.057**	**0.053**
	0.064	smm (2007-12-27)		82.3	0.645	0.101	2.175	0.046

BF – fraction of 9-mer peptides (derived from human and macaque proteome) that are predicted to bind to HLA-B and Mamu alleles
N – number of available experimental measurements used to test prediction algorithms
ACC – % accuracy of classifying peptides into binders and non-binders (Eq. S1)
MCC – Matthews correlation coefficient (Eq. S2)
$\Delta\ln(y)$, σ – average bias and error of predicted binding affinity value IC_{50} (Eq. S6)

Detailed description of each quantity is available in Supplementary Notes 1. HLA-B and Mamu alleles for which prediction algorithms are sufficiently accurate (Supplementary Notes 1) are highlighted with yellow color. The most accurate predictive algorithm is marked with bold font.

Supplementary Table S1: continued

Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles

allele	BF	predictive algorithm	N	ACC	MCC	$\Delta\ln(y)$	σ	$\lvert\Delta\ln(y)/\sigma\rvert$
Mamu-A*11	**0.021**	**ann (2009-09-01)**	367	**91.3**	**0.823**	**-0.034**	**1.532**	**0.022**
	0.018	ann (2007-12-27)		90.5	0.806	-0.056	1.685	0.033
	0.044	smm (2009-09-01)		89.1	0.778	0.147	1.908	0.077
	0.054	smm (2007-12-27)		88.6	0.767	0.129	1.864	0.069
Mamu-B*17	0.004	ann (2009-09-01)	589	88.6	0.763	0.104	1.101	0.095
	0.001	ann (2007-12-27)		71.3	0.403	1.141	2.397	0.476
	0.005	**smm (2009-09-01)**		**83.5**	**0.658**	**0.069**	**1.392**	**0.050**
	0.003	smm (2007-12-27)		69.8	0.356	0.350	1.862	0.188

BF – fraction of 9-mer peptides (derived from human and macaque proteome) that are predicted to bind to HLA-B and Mamu alleles
N – number of available experimental measurements used to test prediction algorithms
ACC – % accuracy of classifying peptides into binders and non-binders (Eq. S1)
MCC – Matthews correlation coefficient (Eq. S2)
$\Delta\ln(y)$, σ – average bias and error of predicted binding affinity value IC_{50} (Eq. S6)

Detailed description of each quantity is available in Supplementary Notes 1. HLA-B and Mamu alleles for which prediction algorithms are sufficiently accurate (Supplementary Notes 1) are highlighted with yellow color. The most accurate predictive algorithm is marked with bold font.

Supplementary Table S2:

Alleles with significant association of HIV control or progression:

allele	OR	95% CI	p value
HLA-B*0702	1.90	[1.41 , 2.56]	1×10^{-3}
HLA-B*2705	0.45	[0.30 , 0.67]	3×10^{-3}
HLA-B*3501	1.95	[1.38 , 2.76]	7×10^{-3}
HLA-B*5701	0.28	[0.19 , 0.42]	1×10^{-8}
HLA-B*5703	0.13	[0.07 , 0.26]	2×10^{-7}

OR (see Fig. 3 caption) – ratio of odds of progressing to high viral loads to controlling HIV to less than 2,000 copies of the virus/ml plasma when expressing a particular HLA allele. The results are corrected for the effects of HLA-B*0702, HLA-B*3501, HLA-B*2705 and HLA-B*5701.

95% CI – 95% confidence interval for OR

Supplementary Table S3:

Parameters of Model shown in Fig 2:

Parameter	Symbol	Value	Units	References
Initial target cell concentration	$I^t(t=0)$	3 x 10^4	cells ml^{-1}	[35]
Maximum virus replication	k_v	2000	virions $(cell\ day)^{-1}$	[41,42]
Virus clearance	k_c	20	day^{-1}	[41]
Mutation rate	k_m	2.2 x 10^{-5}	mutations $(base\ cycle)^{-1}$	[43]
Target cell production	k_b	1000	$(cell\ day)^{-1}$	
Target cell death	k_d	0.1	day^{-1}	
Target cell infection	k_t	6.5 x 10^{-7}	ml $(virus\ day)^{-1}$	[36]
Infected cell death	k_d'	0.15	day^{-1}	[44]
Presentation of pMHC on infected cells, APCs	k_s, k_s'	10	day^{-1}	
Peptide off-rate	k_o, k_o'	1	day^{-1}	[45]
Activated $CD8^+$ expansion	k_p	3	day^{-1}	[46]
Rate of CTL activation/killing	k_a, k_k	4 x 10^{-6}	ml $(cell\ day)^{-1}$	
Memory cell activation	k_{ra}	8 x 10^{-6}	ml $(cell\ day)^{-1}$	
Effector CD8 cell death	k_{dt}	0.5	day^{-1}	[47]
Differentiation of effector to memory cell	k_m	0.008	day^{-1}	
Memory cell death	k_{dm}	0.015	day^{-1}	[48]

Supplementary Table S4:

Parameters of simplified model shown in Figure S7:

Parameter	Symbol	Value	Units
Target cell concentration	I^t	10^4	cells ml^{-1}
Infected cell death	k_d	0.1	day^{-1}
Presentation of pMHC on infected cells, APCs	k_s, k_s'	800	day^{-1}
Peptide off-rate	k_o, k_o'	40	day^{-1}
Activated CTL expansion	k_p	0.2	day^{-1}
Rate of CTL activation/ infected cell killing	k_a / k_k	6×10^{-5}	ml $(cell\ day)^{-1}$

Parameters not listed are the same as in Table S3.

Supplementary Figure S1:

a) HLA-B*5701 ann (2009-09-01) b) HLA-B*2705 ann (2009-09-01)

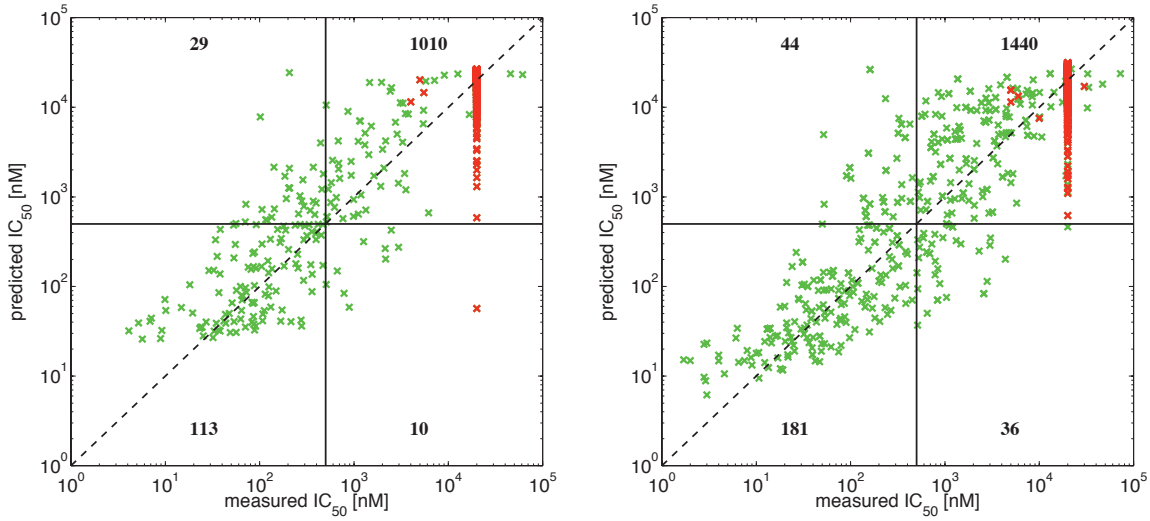

Figure Legend S1: Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*5701 allele (a) and HLA-B*2705 (b). For both alleles the best predictive algorithm (Table S1) was used. Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC_{50} is larger than that corresponding to it's value on the abscissa. Solid lines represent threshold value 500nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (*ACC*) and Matthews correlation coefficient (*MCC*).

Supplementary Figure S2:

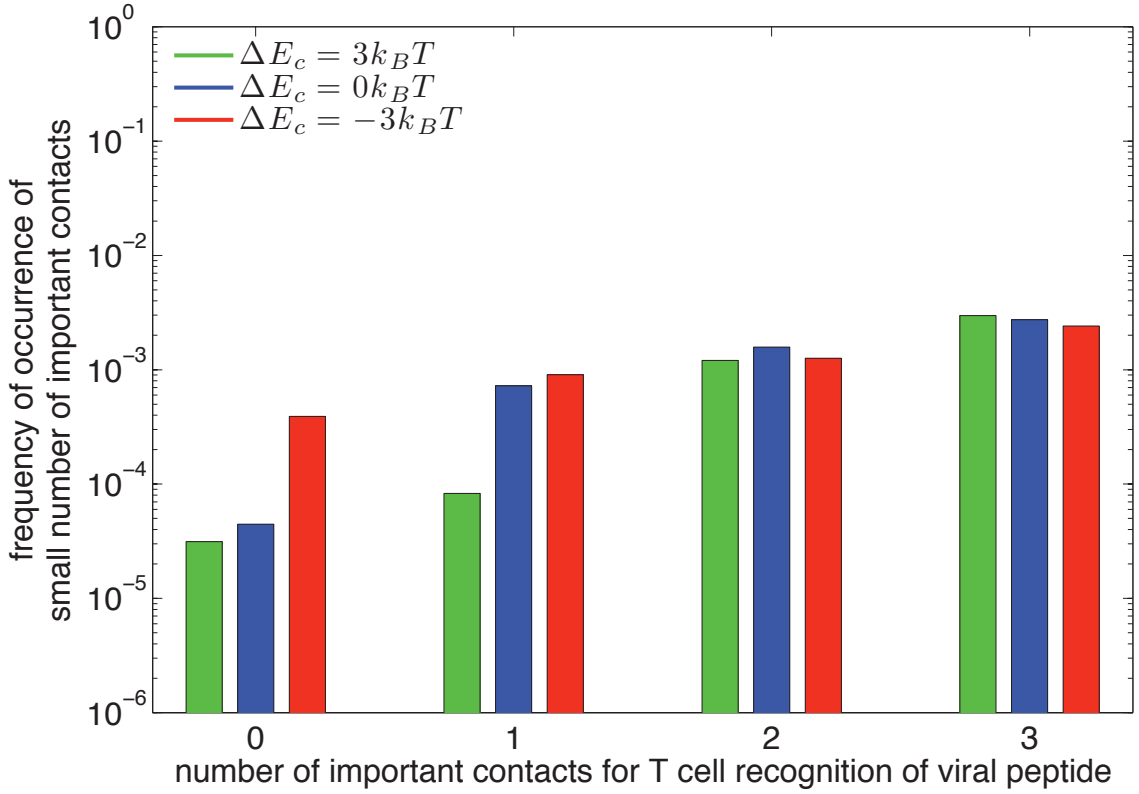

Figure Legend S2: Distribution of the number of important contacts for TCR recognition of antigenic peptides is invariant to variations in interaction free energy (E_c) between TCRs and MHC as long as these interactions are not too strong or too weak. As shown previously[9,10] too strong or weak TCR-MHC interactions result with high probability in T cell deletion in the thymus, because such T cells are negatively selected or not positively selected, respectively. $\Delta E_c = 0$ corresponds to results in the main text, while stronger (weaker) binding is denoted with $\Delta E_c < 0$ ($\Delta E_c > 0$). TCRs were selected against *1000* self peptides. In these calculations we varied the TCR-HLA interaction (E_c) by actually varying the difference between this quantity and the negative selection threshold (E_n). Therefore, this study is equivalent to leaving the value of TCR-HLA interactions the same and varying the value of the binding threshold for negative selection. In this case red (green) bars corresponds to weaker (stronger) binding threshold for negative selection.

Supplementary Figure S3:

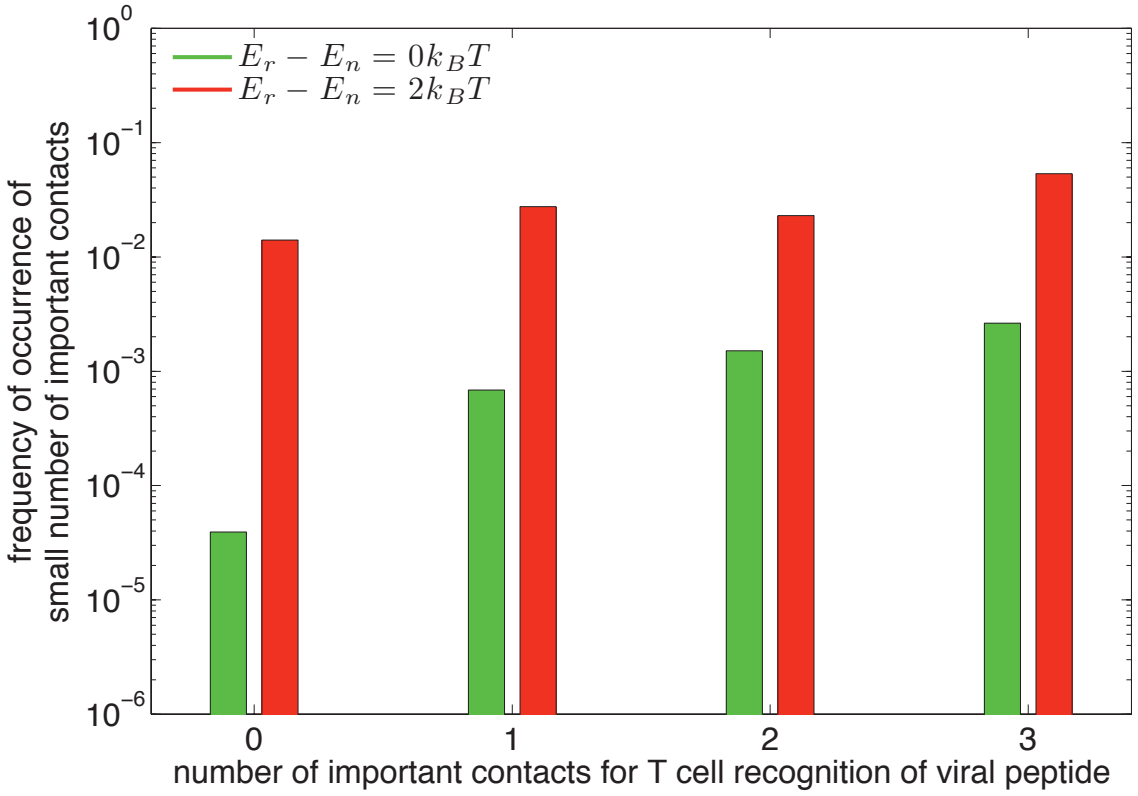

Figure Legend S3: Weaker binding free energy threshold for antigen recognition than the negative selection threshold in the thymus results in more cross-reactive TCRs. Histogram of important contacts for TCR recognition of antigenic peptides for different binding thresholds of antigen recognition for TCRs selected against *1000* self peptides. The green histogram corresponds to the recognition threshold, E_r, being equal to the negative selection threshold (E_n). When threshold for recognition is weak (red histogram), most TCRs are very cross-reactive, because single amino acid mutation on the antigenic peptide is not enough to make the binding interaction free energy weaker then recognition threshold. Experimental evidence suggests that the negative selection threshold in the thymus is the same as recognition threshold in the periphery[12].

Supplementary Figure S4:

a) HLA-B*0801 smm (2009-09-01) b) HLA-B*0801 smm (2007-12-27)

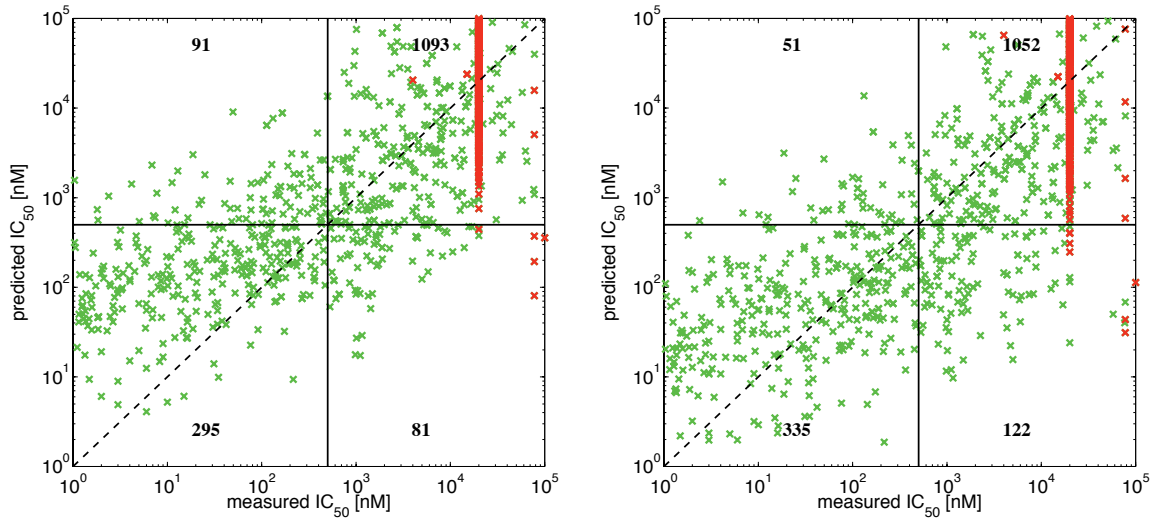

Figure Legend S4: The predictive algorithms for HLA-B*0801 are not very accurate (see also Table S1). Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*0801 allele for two predictive algorithms: smm (2009-09-01) on left and smm (2007-12-27) on right. Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC_{50} is larger than that corresponding to it's value on the abscissa. Solid lines represent threshold value 500nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. Newer algorithm (a) on average tends to overestimate IC_{50} value, which results in predicting fewer peptide binders. Older algorithm (b) on average tends to underestimate IC_{50} value, which results in predicting more peptide binders. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (*ACC*) and Matthews correlation coefficient (*MCC*).

Supplementary Figure S5:

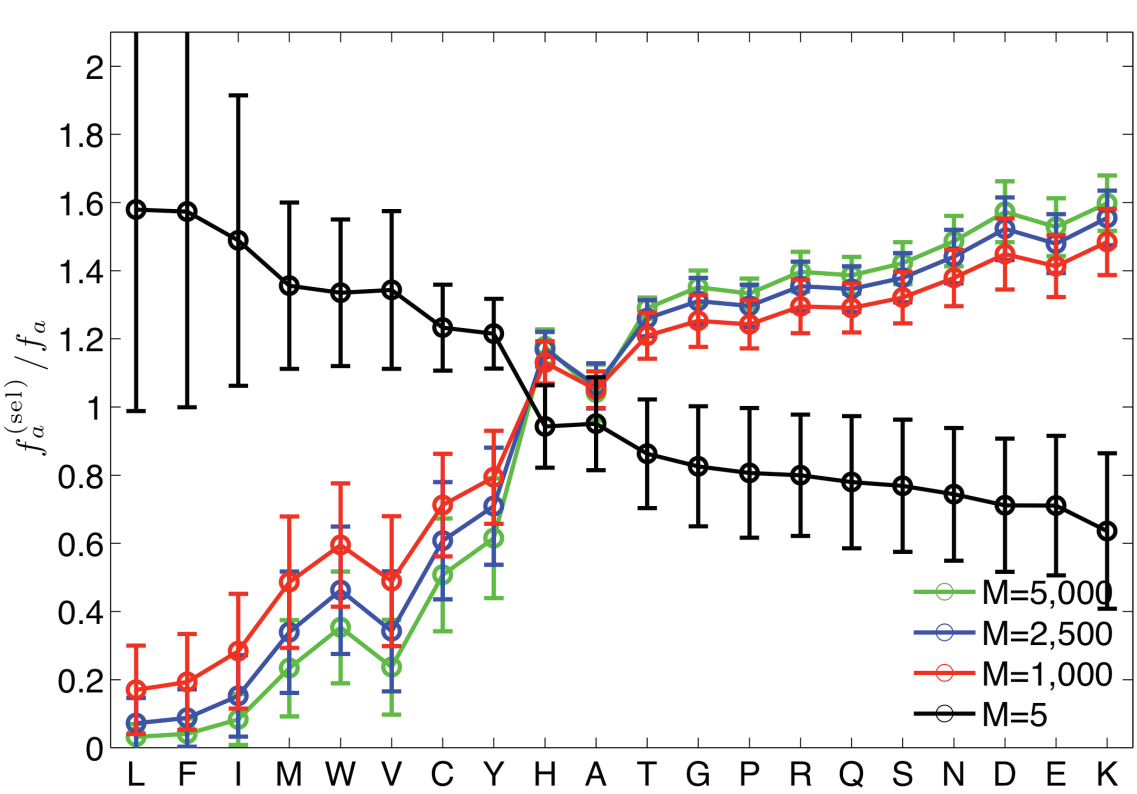

Figure Legend S5: Selection against a greater diversity of peptides (*M*) in the thymus results in selected TCRs with peptide contact residues that are more enriched in amino acids that interact weakly with other amino acids. The ordinate is the ratio of the frequencies of occurrence of an amino acid in the peptide contact residues of selected TCRs ($f_a^{(\mathrm{sel})}$) to preselection TCRs (f_a). The abscissa is a list of amino acids ordered according to the average interaction free energy (as per the MJ interaction potential) with which it interacts with all other amino acids (L – the strongest, K – the weakest). This qualitative result is robust to changes in the interaction potential as can be deduced from analytical and computational results noted in [9,10].

Supplementary Figure S6:

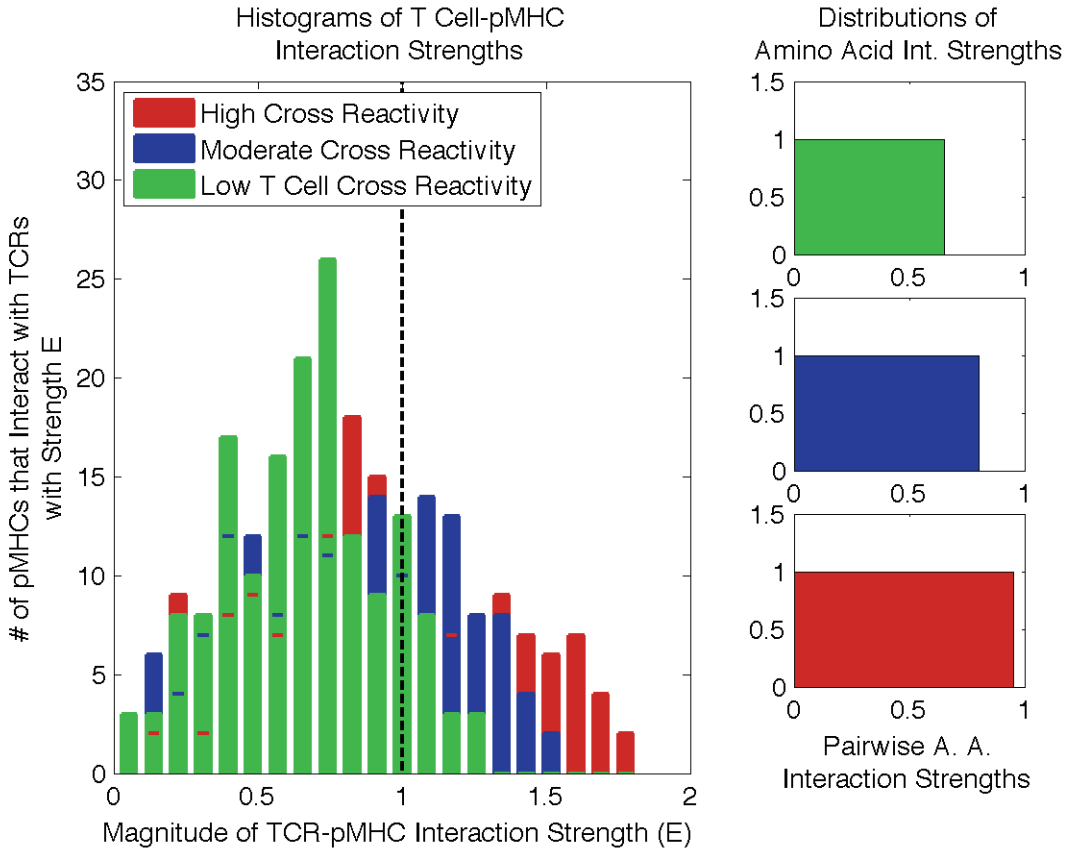

Figure Legend S6: Random energy-like model for generating $\sigma_{i,j}$, the matrix describing recognition of pMHCs by $CD8^+$ T cells. The degree of cross-reactivity in the simulation depends on the uniform distribution from which interaction strengths between individual epitope residues and the TCRs are randomly selected (right). A higher upper limit of the pairwise distribution corresponds to a higher mean and broader distribution of the overall (summed) TCR-pMHC interaction strengths (left). As recognition is considered to occur above a threshold, a broader distribution results in more frequent recognition of pMHCs by T cells in the model, and thus higher cross-reactivity.

Supplementary Figure S7:

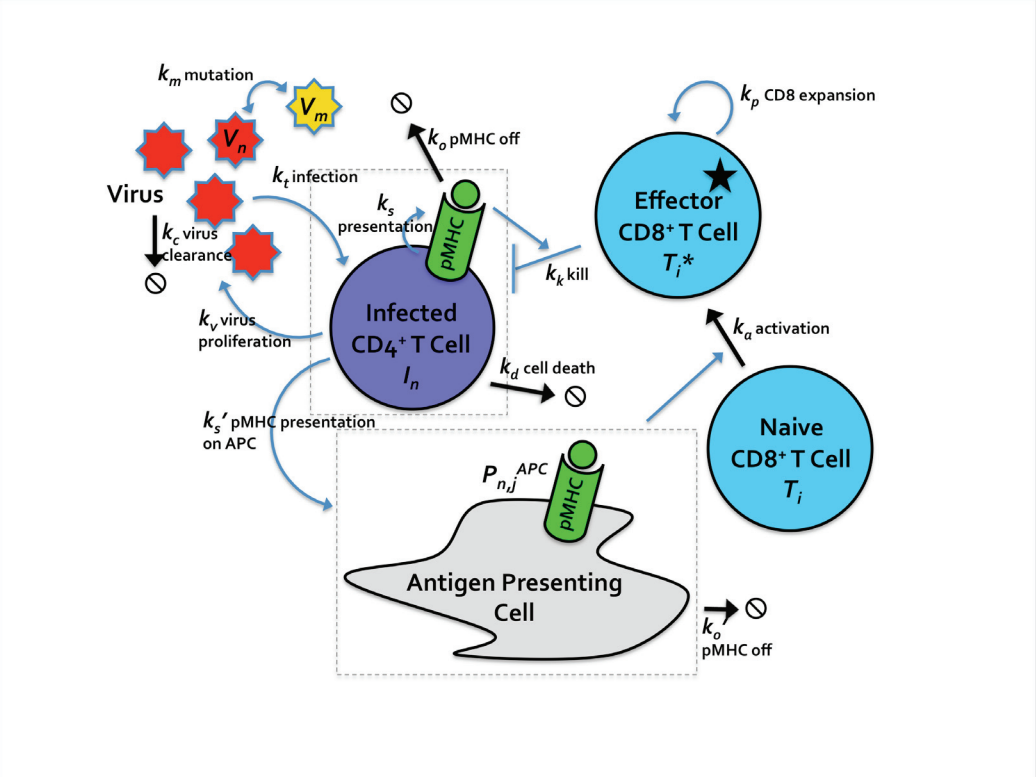

Figure Legend S7: Schematic of a model for host-pathogen dynamics that is simpler than that shown in Fig. 2 (henceforth termed, simplified model). The virus mutates, infects target CD4⁺ T cells, and is cleared. Infected CD4⁺ T cells produce more free virus, and die. Infected cells present viral peptides in complex with HLA molecules for a period (until peptides unbind from HLA). Activated (effector) CD8⁺ T cells produced by recognition of viral epitopes on APCs proliferate and kill infected cells bearing their cognate peptide-HLA complex

Supplementary Figure S8:

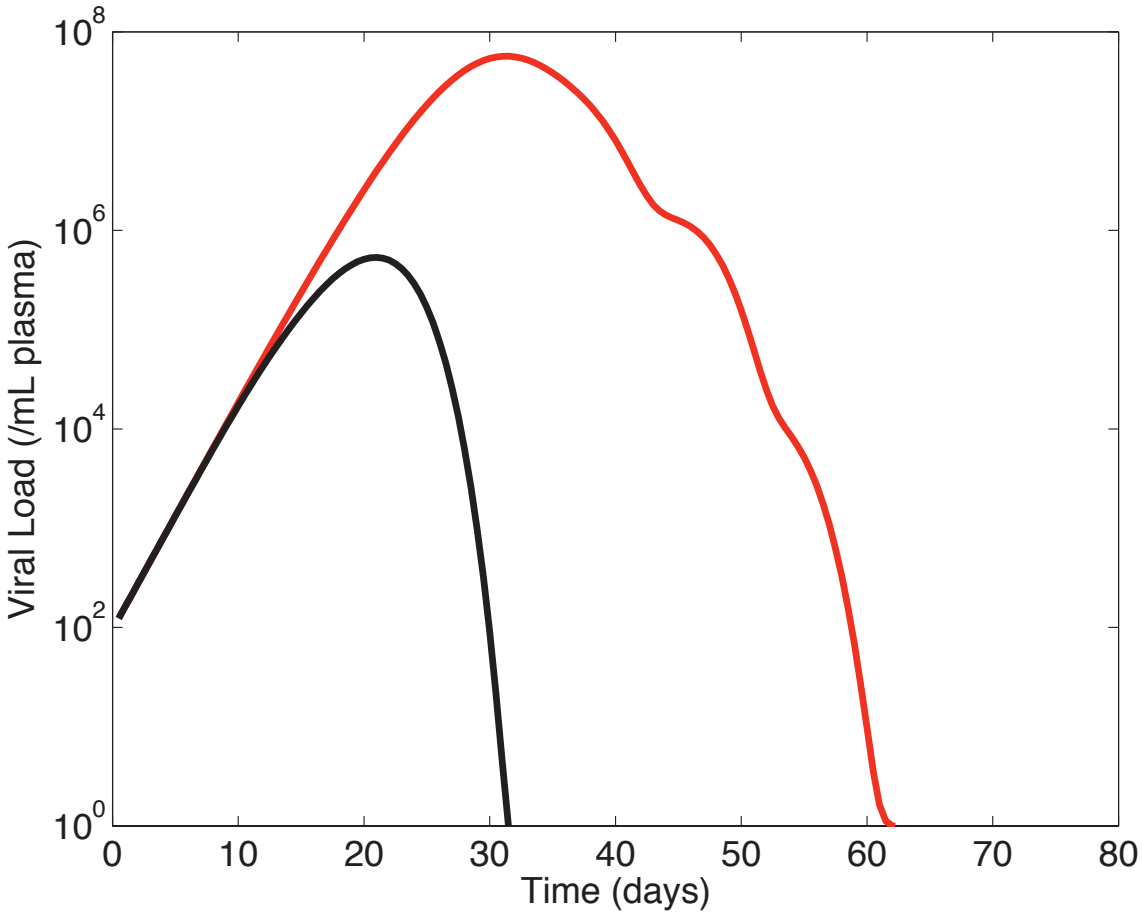

Figure Legend S8: Simulation results using the simplified model. HIV viral loads versus time for different cross-reactivities (CR) of the $CD8^+$ T cell repertoire, corresponding to the model in Fig. S7. Black curve: highly cross-reactive case. Red curve: lower cross-reactivity. Each curve is averaged over 500 simulations (each simulation represents a person).

Supplementary Figure S9:

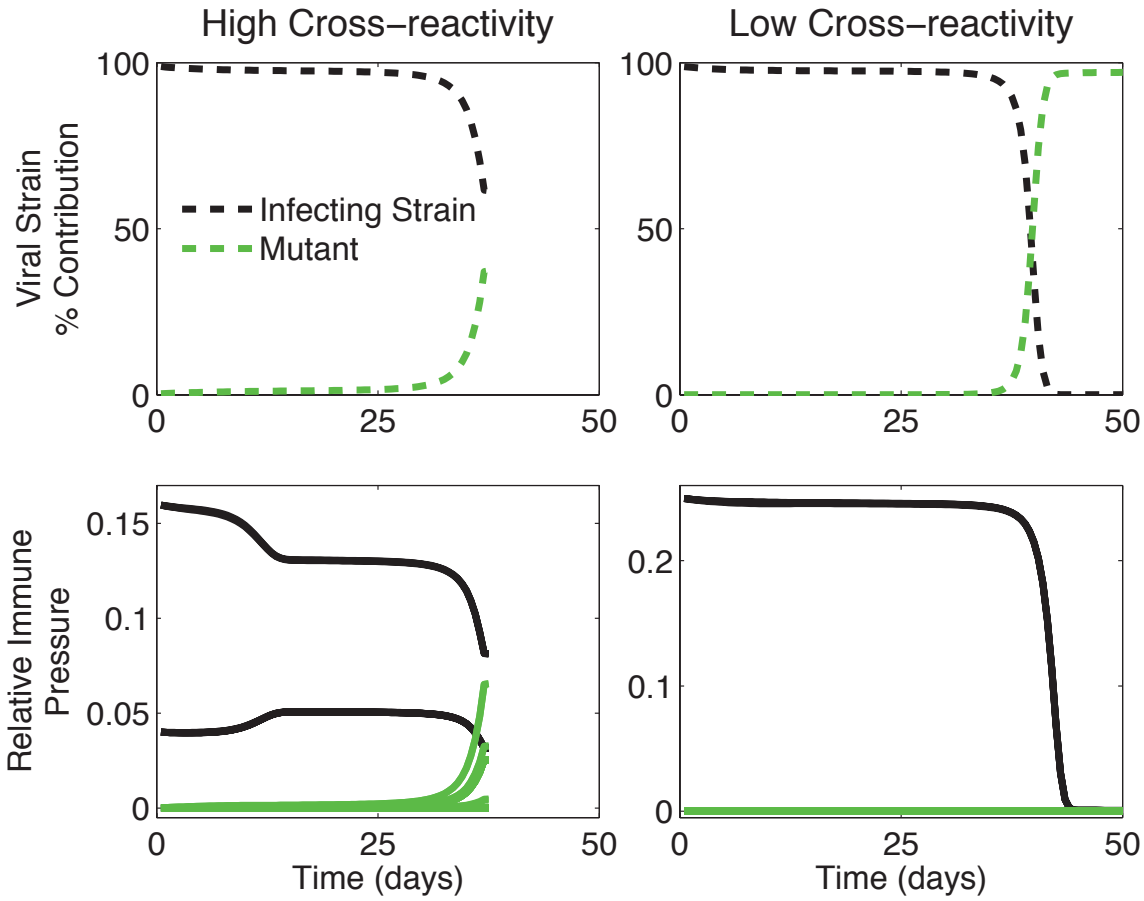

Figure Legend S9: As in Fig. 2c, but for the simplified model (schematic in Fig S7). When more clones recognize the infecting and emerging strains (left, bottom), the emerging mutant strain (green) is kept in check (left, top). However, when cross-reactivity is low, the likelihood that the mutant strain goes unrecognized is higher (bottom, right), and the mutant strain achieves a large percent contribution of the total virus population (top, right).

Supplementary Figure S10:

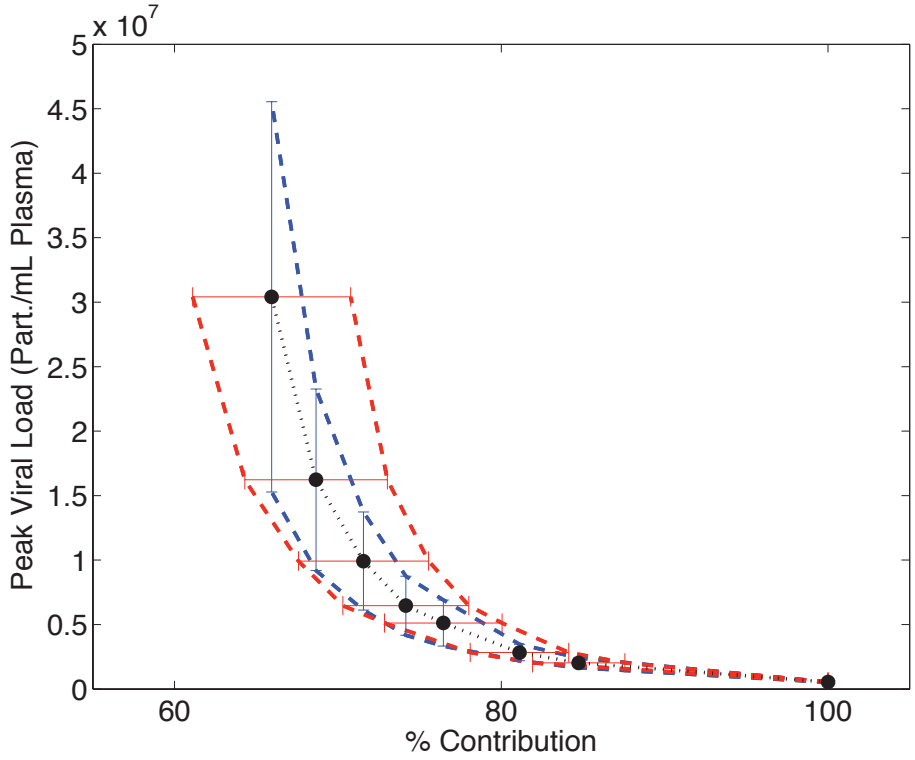

Figure Legend S10: Anticorrelation of simulated peak viral loads with percent contribution of the dominant epitope to the total CTL response for the model in Fig. S7. Percent contribution is calculated as the number of activated CTLs recognizing the immunodominant epitope over the total number of activated CTLs in the simulation. The immunodominant epitope is defined as the epitope recognized by the largest number of CTL clones. Lower percent contributions are achieved when the $CD8^+$ T cell repertoire is less cross-reactive, which also correlates with higher viral loads, as found experimentally by Altfeld and coworkers[30]. The black points and bars correspond to the average and standard deviation of 500 simulations for each level of T cell cross-reactivity, with the level of cross-reactivity increasing from left (probability of .28 that a given epitope is recognized by a particular CTL) to right (probability 1). Varying other parameters in the model, including peptide presentation rate, does not capture this behavior (Fig. S16).

Supplementary Figure S11:

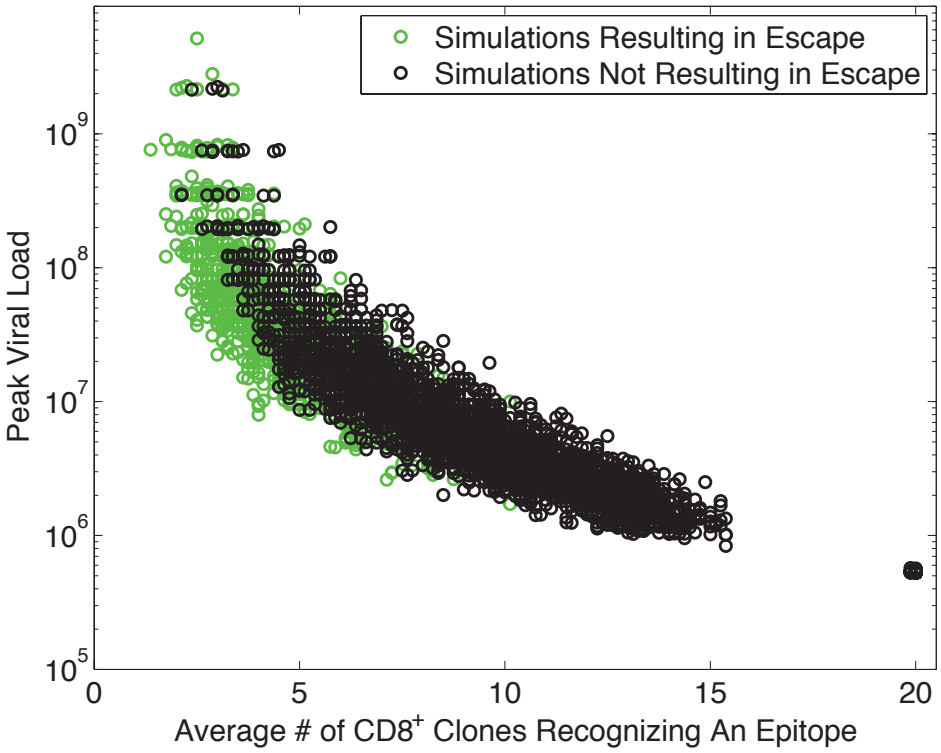

Figure Legend S11: Peak viral load versus average number of $CD8^+$ clones recognizing a pMHC in each simulation in the simplified model (schematic in Fig. S7). "Escape" is taken to mean that the population size of a mutant viral strain has become larger than that of the infecting strain at some point during the simulation time (0 to 80 days). As exemplified in Fig. 2c, the smaller the number of clones recognizing each pMHC (corresponding to lower cross-reactivity), the higher the chance of escape.

Supplementary Figure S12:

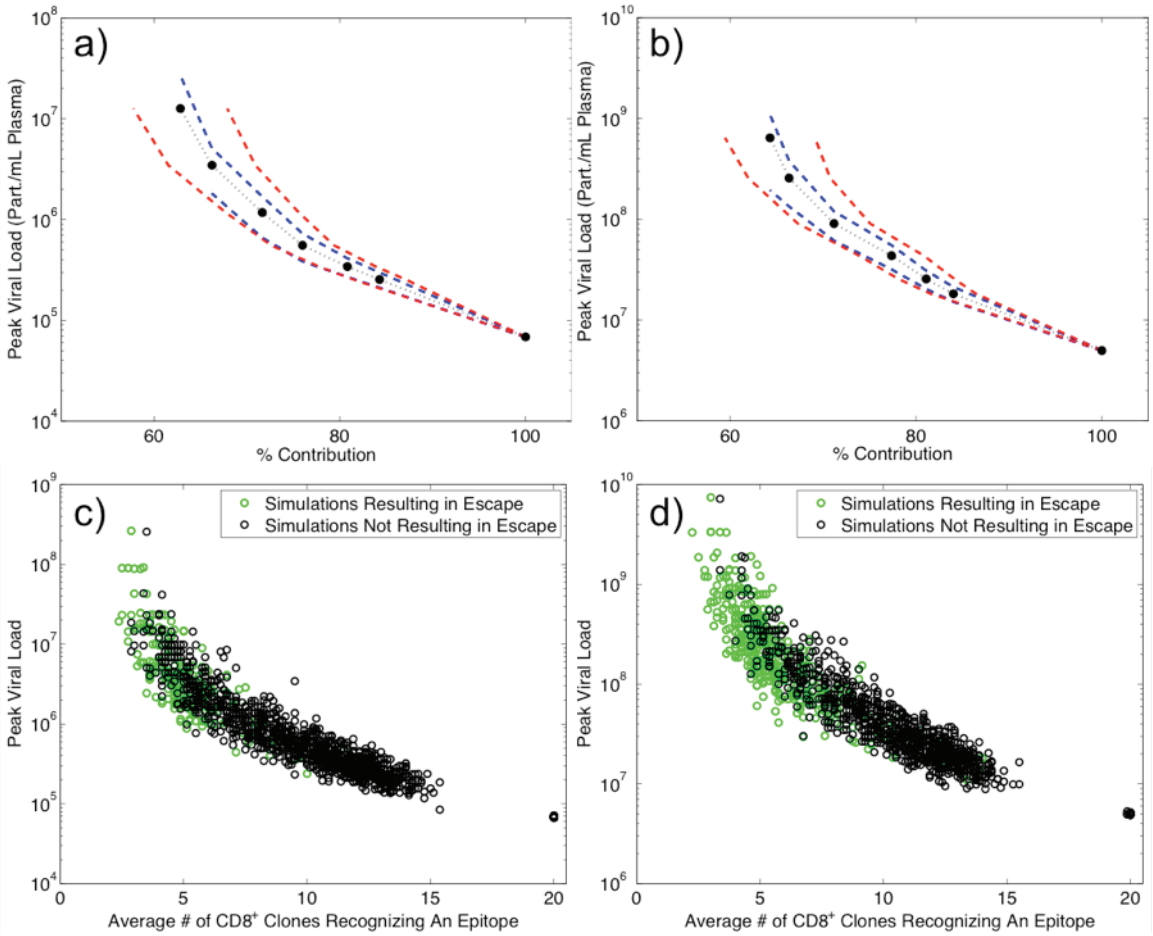

Figure Legend S12: Insensitivity of qualitative results to changes in $CD8^+$ T cell activation rate for the simplified model (schematic in Fig. S7). Left panels show simulation results with varying cross reactivity for activation rate $10 * k_a$ (k_a given in Table S4), while right panels show results for $k_a/10$. Insensitivity of qualitative results to parameter variation was found for other rate constants in the model also (not shown).

Supplementary Figure S13:

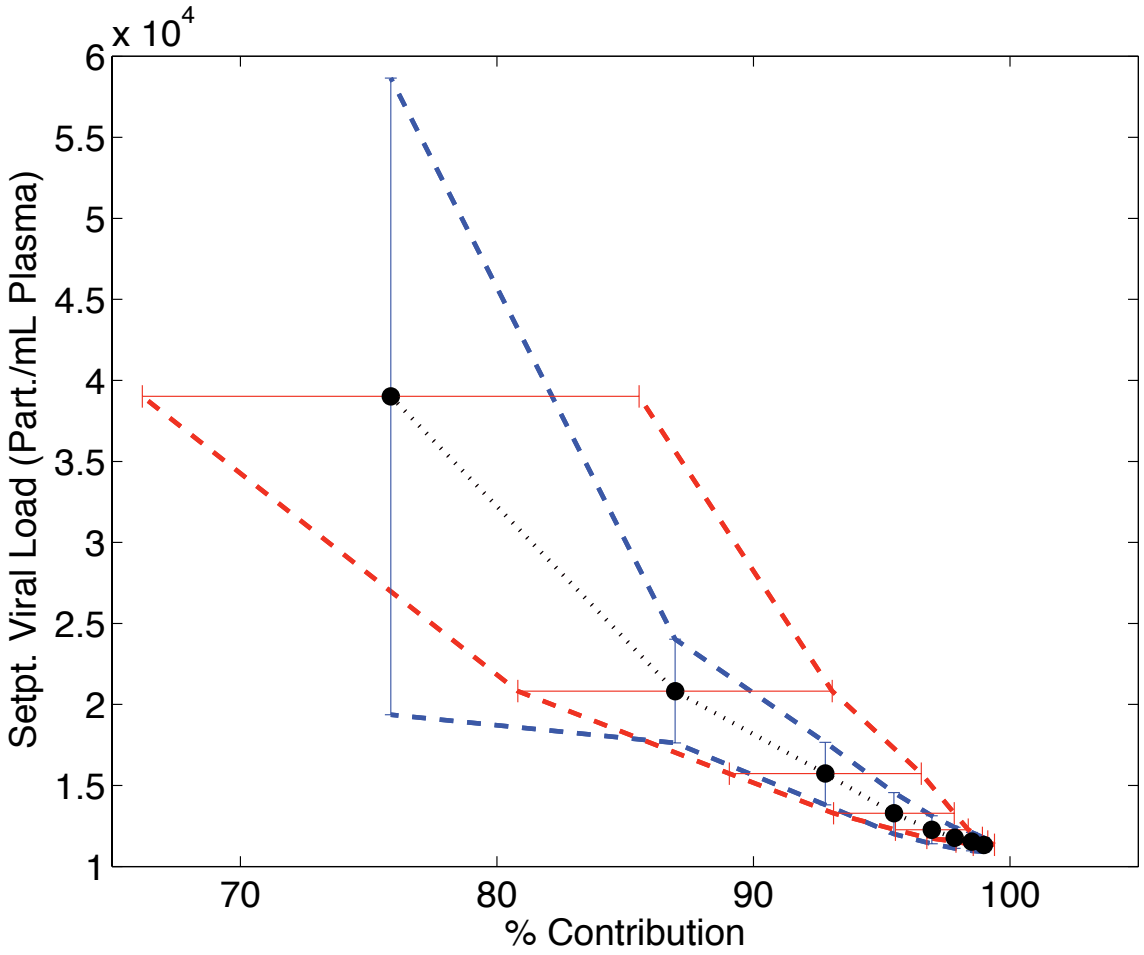

Figure Legend S13: Anticorrelation of simulated viral loads with percent contribution of the dominant epitope to the total CTL response, as in Fig. S10, but corresponding to the model discussed in the main text (Fig. 2a). Viral load and % contribution were calculated at day 200 in the simulations, to approximate viral load setpoint. Both models give qualitatively similar results. Thus, the result that a cross-reactive repertoire results in low viral loads and high % contribution of responses to the dominant epitope is insensitive to the choice of dynamical model.

Supplementary Figure S14:

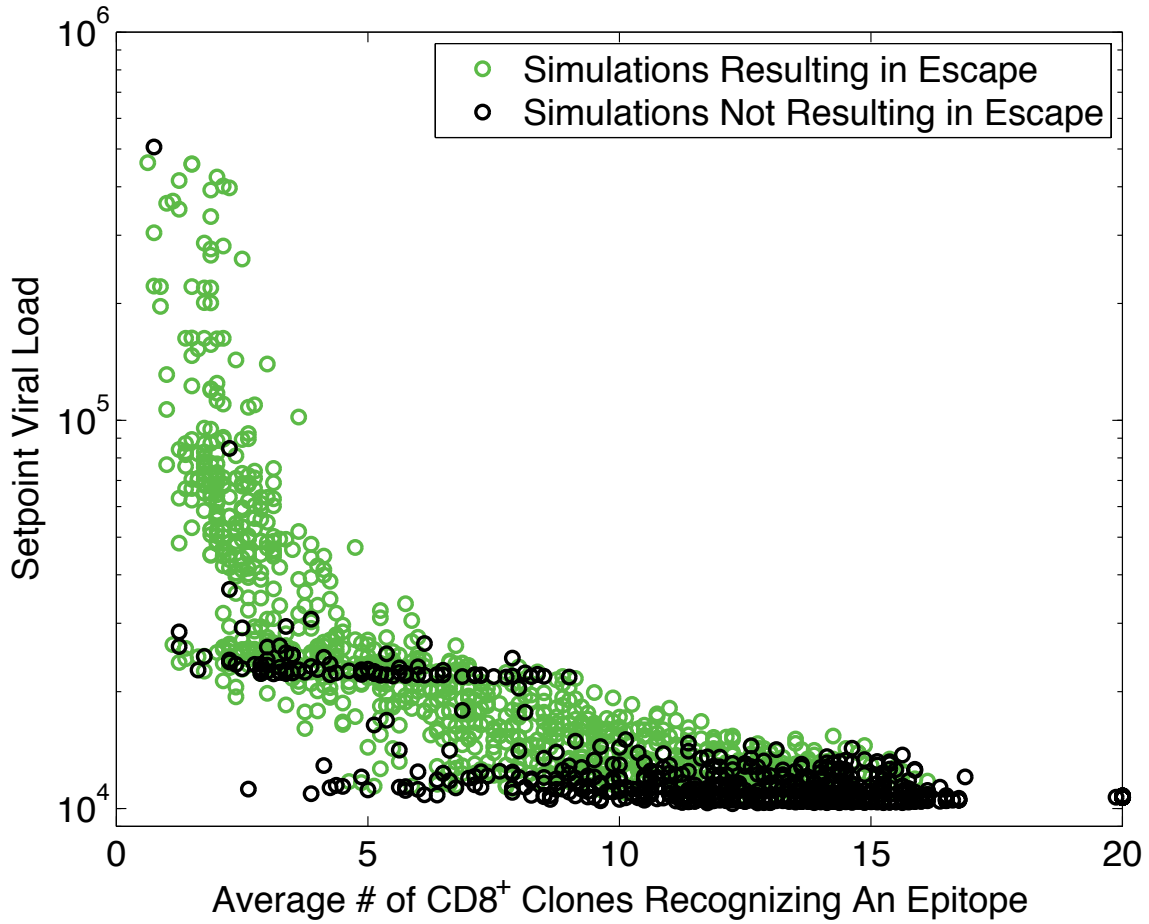

Figure Legend S14: As in Fig. S11, but for the model described in the main text (schematic in Fig. 2a). As the number of clones recognizing a pMHC increases, the setpoint viral load and probability of escape decrease.

Supplementary Figure S15:

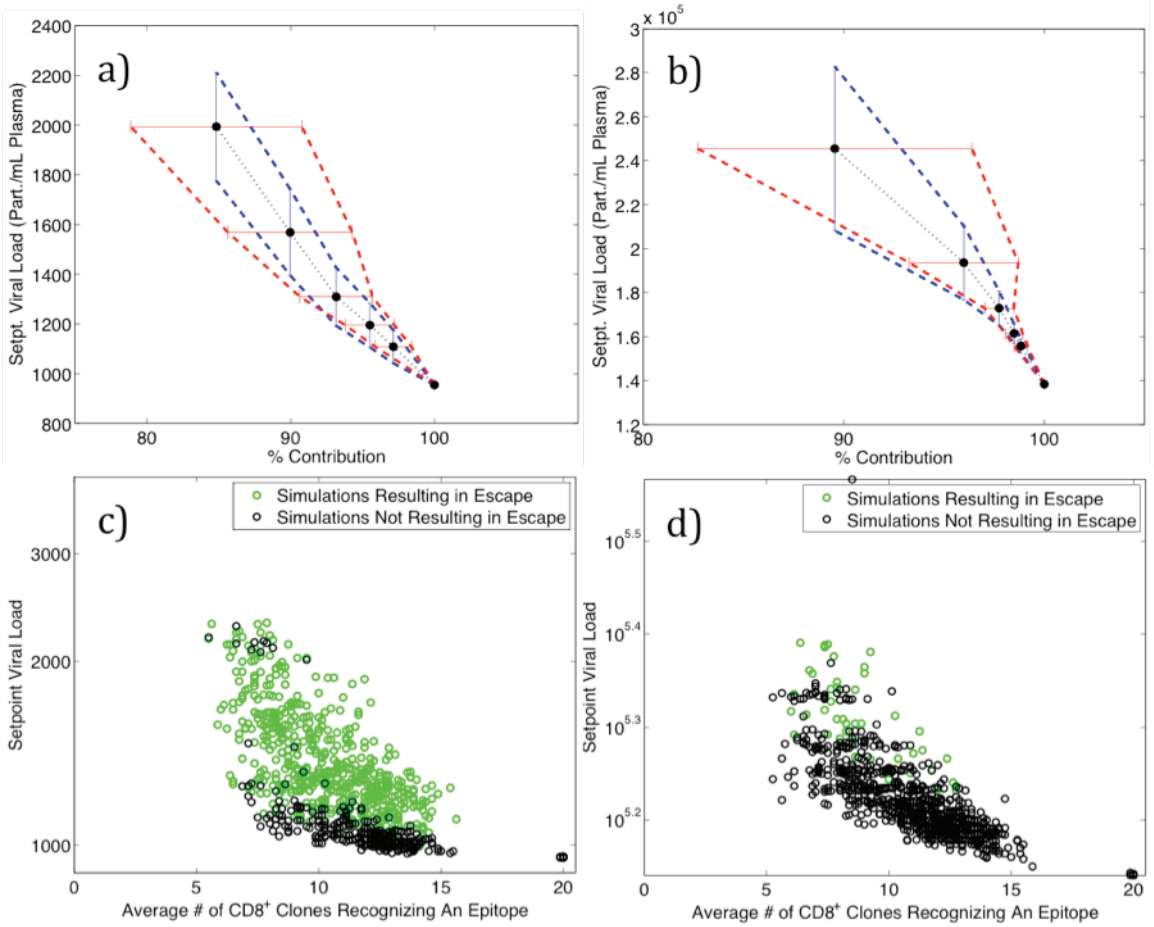

Figure Legend S15: Insensitivity of qualitative results to changes in $CD8^+$ T cell activation rate for the model described in the main text (schematic in Fig. 2a). Left panels show simulation results with varying cross reactivity for activation rate $10 * k_a$ (k_a given in Table S3), while right panels show results for $k_a/10$. The setpoint viremia level depends strongly on k_a, but the qualitative correlation between viral load and % contribution of the immunodominant epitope (a and b) and the number of clones targeting a given pMHC (c and d) is the same. The same insensitivity of qualitative results to parameter variation was found for other rate constants in the model also (not shown). Note that for a lower activation rate (right panels), the probability of escape is reduced (fewer green points), because of the lower overall immune pressure exerted by the same number of T cells.

Supplementary Figure S16:

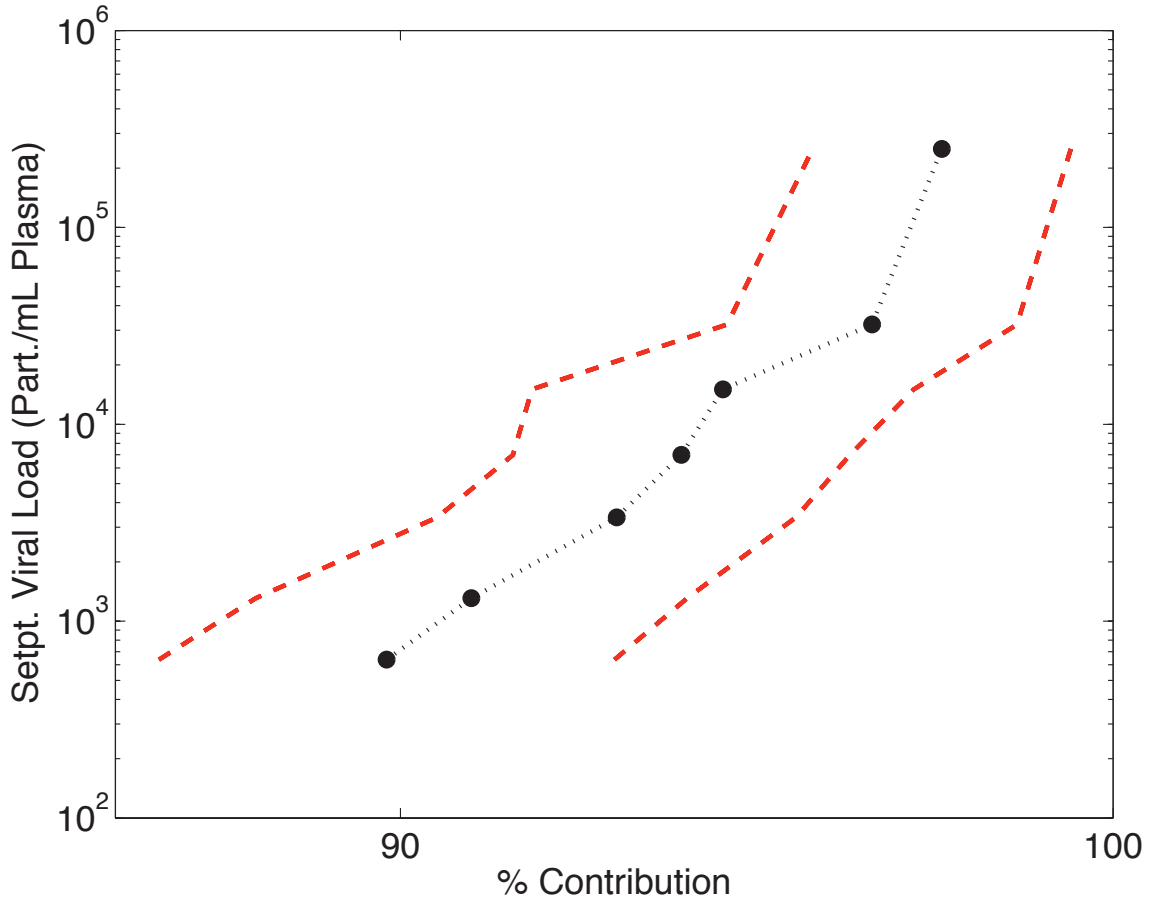

Figure Legend S16: Setpoint viral load versus % contribution of immunodominant epitope, as in Figs. S10 and S13, but where the rate of peptide presentation k_s (not cross-reactivity), is varied. Points correspond to k_s values of 200, 100, 40, 20, 10, 5, and 1 (day^{-1}), with faster presentation of pMHC corresponding to reduced peak viral loads. One potential effect of B57 binding fewer peptides is that the cell-surface concentration of immunogenic peptides could increase, because competition with other peptides for binding to MHC would be reduced. This would then be an additional mechanism for control of viral load. Increasing k_s has the effect of raising cell-surface concentration of pMHC and reducing viral load. As the figure shows, varying only this parameter leads to correlation of high % contribution with high viral loads, in contrast to the result of Altfeld, coworkers[30]. Variation of other rate constants in the model gave similar results or had no effect on % contribution (not shown). Therefore, only varying the cross-reactivity recapitulates the experimental results.

The issue of peptide presentation could be important if HLA molecules like HLA-B*5701 presented far fewer HIV epitopes and so, due to less competition, these epitopes were presented faster, and hence, in greater amounts. We have used the predictive algorithms and the published HIV proteome (HXB2) to estimate the number of HIV epitopes that can bind to the alleles we have identified from our data (Fig. 3) to be associated with control or progression. Approximately 40 peptides can bind to HLA-B*5701 and HLA-B*0702 and approximately 60 peptides can bind to HLA-B*2705 and HLA-B*3501. Thus, the number of HIV peptides that can bind to these alleles does not correlate with disease outcome.

Supplementary Figure S17:

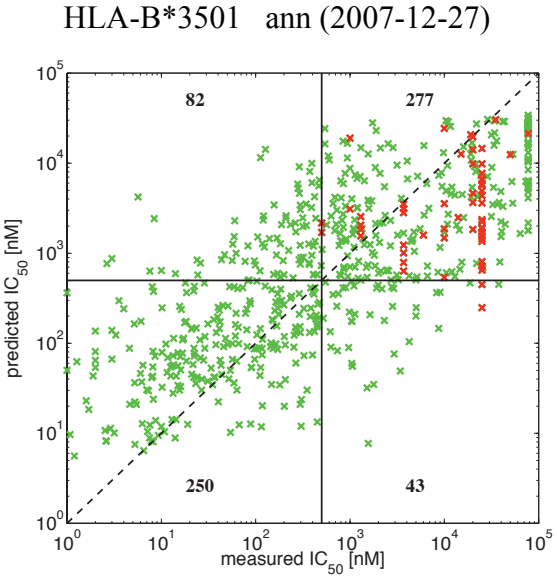

Figure Legend S17: The predictive algorithms for HLA-B*3501 are less accurate than that fo HLA-B*5701, HLA-B*0702, and HLA-B*2705 (see also Table S1). Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*3501 allele for the most accurate predictive algorithm ann (2007-12-27). Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC_{50} is larger than that corresponding to it's value on the abscissa. Solid lines represent threshold value 500nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. The algorithm on average tends to overestimate IC_{50} value, which results in predicting a smaller peptide binding fraction than reality. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (*ACC*) and Matthews correlation coefficient (*MCC*).

Supplementary Methods: Host-pathogen interaction dynamics for simplified model

The following dynamical model is similar to that in Fig. 2a, but is without the effects of target cell limitation, finite CTL expansion, and $CD8^+$ memory. Similar models have been studied previously[49]. The following equations describe the model (schematic in Fig. S7):

$$\frac{dV_n}{dt} = k_v^n I_n - k_c V_n + k_m \sum_{n:m} (V_m - V_n) \quad \text{(S7)}$$

$$\frac{dI_n}{dt} = k_t V_n I^t - k_d I_n - \sum_i \sum_j \sigma_{i,j} k_k P_{n,j} T_i^* \quad \text{(S8)}$$

$$\frac{dP_{n,j}}{dt} = k_s^j I_n - k_0^j P_{n,j} - \frac{dI_n^{(kill)}}{dt} \frac{P_{n,j}}{I_n} \quad \text{(S9)}$$

$$\frac{dP_{n,j}^{APC}}{dt} = k'^j_s I_n - k'^j_0 P_{n,j}^{APC} \quad \text{(S10)}$$

$$\frac{dT_i}{dt} = -k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC} \quad \text{(S11)}$$

$$\frac{dT_i^*}{dt} = k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC} + k_p T_i^* \quad \text{(S12)}$$

Rate parameters for the more complex model in the main text (Fig. 2) and the model above are given in Tables S3 and S4, respectively. Rate constants governing virus and $CD4^+$ dynamics are generally adopted from the literature. Approximate rate constants for virus and $CD4^+$ cell turnover are available from studies in which patient viral loads were perturbed by antiretroviral treatment or plasma apheresis, and the data were fit by dynamical models[41,50,51]. Predicted rate constants for infected $CD4^+$ cell death range from about 0.1 to 1 (day^{-1})[44]. This rate constant accounts for cell death due to virus cytotoxicity as well as clearance by effector CTLs and antibodies, and thus is considered an upper bound for k_d in our model, which describes infected cell death by means other than CTL killing. Estimates for the percentage of infected cell death attributable to the CTL response range from 10% to 90%[18,52]. Constants for reactions involving $CD8^+$ cells are chosen to give realistic peak and setpoint (in the model described in the main text only) viral loads. The mutation rate from the literature in units of mutations $(base\ cycle)^{-1}$

is converted to ~.22/(L*M) mutations (amino acid day)$^{-1}$ using an estimate of 1 day for a replication cycle[53] and ~10^4 base pairs for the size of the virus. In the chronic infection model, the number of cell divisions (D) is taken to be 8^{20}.

If the parameters in the model are chosen such that the virus is able to take hold and expand[54], the qualitative results related to the effects of cross-reactivity are insensitive to the choice of rate constant parameters. This is demonstrated in Fig. S12 for 100-fold variation of the rate constant governing T cell activation, and results were found to be similarly insensitive to variations in the other rate constants (data not shown).

Supplementary Discussion 1: T cells restricted by HLA-B*5701 encounter a smaller diversity of TCR contact residues in the thymus.

Our study showed that the protective allele HLA-B*5701 binds fewer peptides derived from human proteome compared to other alleles (Table S1). Even if the reason why HLA-B*5701 molecules bind fewer self peptides was due to greater restrictions in the tolerance to different amino acids at the anchor residues only, HLA-B*5701 molecules would present a smaller diversity of TCR contact residues in the thymus. This is because the number of self peptides presented in the thymus is much smaller than all possible sequences of TCR contact residues derived from the human proteome. Thus, the probability that any HLA allele presents peptides derived from different parts of the proteome with identical TCR contact residues constrained by the same anchor residues is small. Therefore, since HLA-B*5701 presents fewer self-peptides, T cells restricted by this allele will encounter a smaller diversity of TCR contact residues during development in the thymus.

Supplementary Discussion 2: Additional tests of predictions from the thymic selection model

In our recent work on thymic selection and T cell repertoire development[9,10], we constructed a coarse-grained model that was not quantitative, but yielded qualitative insights that could be directly tested against experiments. Our computational and theoretical studies predicted that, if developing T cells encounter many self peptides in the thymus, their peptide contact residues would be statistically enriched in amino acids that tend to interact weakly with other amino acids (Fig. S5 and [9,10]). To test this prediction we analyzed available crystal structures of TCR-peptide-MHC complexes and found that amino acids determined by bioinformatic studies to be weakly interacting are indeed enriched in TCR peptide contact residues (a detailed discussion of the analyses of crystal structures and comparisons to the theoretical predictions are provided in [9]).

We then predicted that, because of the preponderance of weakly interacting amino acids in the peptide contact residues of mature TCRs, peptide recognition should be mediated by many weak interactions each of which contributes significantly to the binding affinity. Thus, most point mutations to peptide amino acids would abrogate recognition; i.e., specificity. In contrast, if there is one type of self peptide in the thymus (as in the Kappler-Marrack experiments[7,8]), TCRs with strongly interacting amino acids in the peptide contact residues would survive selection (Fig. S5 and [9]). Antigenic peptide recognition by such TCRs would be due to a few strong interactions mediated by these TCR contact residues. Only mutations at peptide amino acids involved in these strong interactions would abrogate recognition, thus making TCR recognition of peptides cross-reactive to mutations at the other sites. These predictions are also supported directly by calorimetric measurements carried out using T cells derived from mice that express one and many types of self peptides in the thymus[7].

Supplementary Notes 2: supplementary references

41 Ramratnam, B. *et al.* Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet* **354**, 1782-1785, (1999).

42 Ribeiro, R. M. Dynamics of CD4(+) T cells in HIV-1 infection. *Immunol Cell Biol* **85**, 287-294, (2007).

43 Huang, K. J. & Wooley, D. P. A new cell-based assay for measuring the forward mutation rate of HIV-1. *J Virol Methods* **124**, 95-104, (2005).

44 Bonhoeffer, S., Funk, G. A., Gunthard, H. F., Fischer, M. & Muller, V. Glancing behind virus load variation in HIV-1 infection. *Trends Microbiol* **11**, 499-504, (2003).

45 Peter, K., Men, Y., Pantaleo, G., Gander, B. & Corradin, G. Induction of a cytotoxic T-cell response to HIV-1 proteins with short synthetic peptides and human compatible adjuvants. *Vaccine* **19**, 4121-4129, (2001).

46 Murali-Krishna, K. *et al.* Counting antigen-specific CD8 T cells: A reevaluation of bystander activation during viral infection. *Immunity* **8**, 177-187, (1998).

47 De Boer, R. J. *et al.* Recruitment times, proliferation, and apoptosis rates during the CD8(+) T-Cell response to lymphocytic choriomeningitis virus. *Journal of Virology* **75**, 10663-10669, (2001).

48 Ladell, K. *et al.* Central memory CD8(+) T cells appear to have a shorter lifespan and reduced abundance as a function of HIV disease progression. *J. Immunol.* **180**, 7907-7918, (2008).

49 Handel, A. & Antia, R. A simple mathematical model helps to explain the immunodominance of CD8 T cells in influenza A virus infections. *J Virol* **82**, 7768-7772, (2008).

50 Wei, X. P. *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117-122, (1995).

51 Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123-126, (1995).

52 Asquith, B., Edwards, C. T. T., Lipsitch, M. & McLean, A. R. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol* **4**, 583-592, (2006).

53 Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582-1586, (1996).

54 Perelson, A. S. & Nelson, P. W. Mathematical analysis of HIV-1 dynamics in vivo. *Siam Review* **41**, 3-44, (1999).