LETTERS

Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection

Andrej Košmrlj[1,2]*, Elizabeth L. Read[1,3,4]*, Ying Qi[5], Todd M. Allen[1], Marcus Altfeld[1], Steven G. Deeks[6], Florencia Pereyra[1], Mary Carrington[1,5], Bruce D. Walker[1,7] & Arup K. Chakraborty[1,3,4,8]

Without therapy, most people infected with human immunodeficiency virus (HIV) ultimately progress to AIDS. Rare individuals ('elite controllers') maintain very low levels of HIV RNA without therapy, thereby making disease progression and transmission unlikely. Certain HLA class I alleles are markedly enriched in elite controllers, with the highest association observed for *HLA-B57* (ref. 1). Because HLA molecules present viral peptides that activate $CD8^+$ T cells, an immune-mediated mechanism is probably responsible for superior control of HIV. Here we describe how the peptide-binding characteristics of HLA-B57 molecules affect thymic development such that, compared to other HLA-restricted T cells, a larger fraction of the naive repertoire of B57-restricted clones recognizes a viral epitope, and these T cells are more cross-reactive to mutants of targeted epitopes. Our calculations predict that such a T-cell repertoire imposes strong immune pressure on immunodominant HIV epitopes and emergent mutants, thereby promoting efficient control of the virus. Supporting these predictions, in a large cohort of HLA-typed individuals, our experiments show that the relative ability of HLA-B alleles to control HIV correlates with their peptide-binding characteristics that affect thymic development. Our results provide a conceptual framework that unifies diverse empirical observations, and have implications for vaccination strategies.

HIV infection leads to acute high level viraemia, which is subsequently reduced to a set-point viral load. Without therapy, most patients experience a subsequent increase in viral load, and ultimately the development of AIDS. Viraemia levels and time to disease vary widely, and the differences correlate with the expression of different HLA class I molecules (reviewed in ref. 2). Effector $CD8^+$ T cells (CTLs) are implicated in viral control because T-cell antigen receptors (TCRs) on $CD8^+$ T cells recognize complexes of viral peptides and class I HLA molecules presented on the surface of infected cells, and depletion of $CD8^+$ T cells leads to increased viraemia in animal models of HIV infection[3]. We describe a feature of the HLA-B57-restricted $CD8^+$ T-cell repertoire that contributes to enhanced control of viraemia.

Algorithms[4] based on experimental data predict whether a particular peptide will bind to a given HLA molecule[5]. We tested four predictive algorithms against available experimental data on peptide binding to diverse HLA molecules and found that, in most cases, they are highly accurate (Supplementary Fig. 1 and Supplementary Table 1). For example, predictions using the best algorithm for HLA-B*5701 were 97% accurate. Using these algorithms, we computed the fraction of peptides derived from the human proteome[6] that bind to various HLA molecules. Of the $\sim 10^7$ unique peptide sequences, only 70,000 are predicted to bind to HLA-B*5701, and 180,000 bind to HLA-B*0701 (an allele that is not protective against HIV). Essentially identical results were obtained for randomly generated peptides (data not shown). The protective allele in macaques, Mamu-B*17, also binds fewer self peptides than other Mamu molecules for which data are available (Mamu-B*17 binds 4, 6 and 13 times fewer self peptides than Mamu-A*11, Mamu-A*01 and Mamu-A*02, respectively; Supplementary Table 1).

The intrinsic differences in self-peptide binding among HLA molecules are important during T-cell repertoire development. Immature T cells are exposed to diverse host-derived peptide–HLA complexes presented in the thymus. As fewer self peptides are able to bind to HLA-B*5701 (and Mamu-B*17) molecules, a smaller diversity of self-peptide TCR contact sequences will be encountered by HLA-B*5701/Mamu-B*17-restricted T cells in the thymus (Supplementary Discussion 1).

The diversity of self peptides presented in the thymus shapes the characteristics of the mature T-cell repertoire. Experiments[7,8] and theoretical studies[9,10] show that T cells that develop in mice with only one type of peptide in the thymus are more cross-reactive to point mutants of peptide epitopes that they recognize than T cells from mice that express diverse self peptides. Thus, by encountering fewer self peptides during thymic development, HLA-B57-restricted $CD8^+$ T cells should be more cross-reactive to point mutants of targeted viral peptides.

We carried out *in silico* thymic selection experiments to test this hypothesis. We chose an HLA-dependent number of thymic self peptides, each with amino acids of the TCR contact residues picked according to the frequency with which they appear in the human proteome[6,9]. A diverse set of immature $CD8^+$ T cells (thymocytes) was generated by choosing the sequences of their peptide contact residues in the same way, and by varying the TCR–HLA interactions. A thymocyte emerges from the thymus as a mature $CD8^+$ T cell if its TCR binds to at least one self-peptide–major histocompatibility complex (pMHC; human MHC is called HLA) molecule with an affinity that exceeds the positive selection threshold, and does not interact with any pMHC more strongly than the negative selection threshold. Using a computational model[9,10] in the class of 'string models'[11], we assessed the affinity of TCR–self-peptide–HLA complexes (Methods) to determine which T cells survive positive and negative selection, and become a part of the mature repertoire. Our qualitative results are independent of the parameters used to determine these interaction strengths (Supplementary Figs 2 and 3)[9,10].

[1]Ragon Institute of MGH, MIT and Harvard, Boston, Massachusetts 02114, USA. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [5]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA. [6]University of California, San Francisco, California 94110, USA. [7]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. [8]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

The mature T cells that emerged from these *in silico* thymic selection experiments were then computationally challenged by a viral peptide (that is, not seen in the thymus) bound to the same HLA type. T cells that recognize this peptide–HLA complex were obtained by assessing whether the interaction strength exceeded the negative selection threshold (shown to be equal to the recognition threshold in mouse models[12]); qualitative results are invariant if the recognition threshold is not much weaker than that corresponding to negative selection (Supplementary Fig. 3). Cross-reactivity of these T cells was then determined *in silico* by mutating each TCR contact residue of the peptide to the other 19 possibilities. Sites on the viral peptide were called 'important contacts' if half the mutations therein abrogated recognition by T cells that target this epitope. The frequency of the number of important contacts in viral peptides that determine T-cell recognition was obtained by repeating this procedure 1,000 times with different choices of thymocytes and self and foreign peptides.

Our calculations predict that a T-cell repertoire restricted by an HLA molecule such as HLA-B*5701, which presents fewer self peptides in the thymus, has a higher frequency of occurrence of T cells that recognize viral peptides through smaller numbers of important contacts (Fig. 1a). In contrast, the frequency of occurrence of T cells

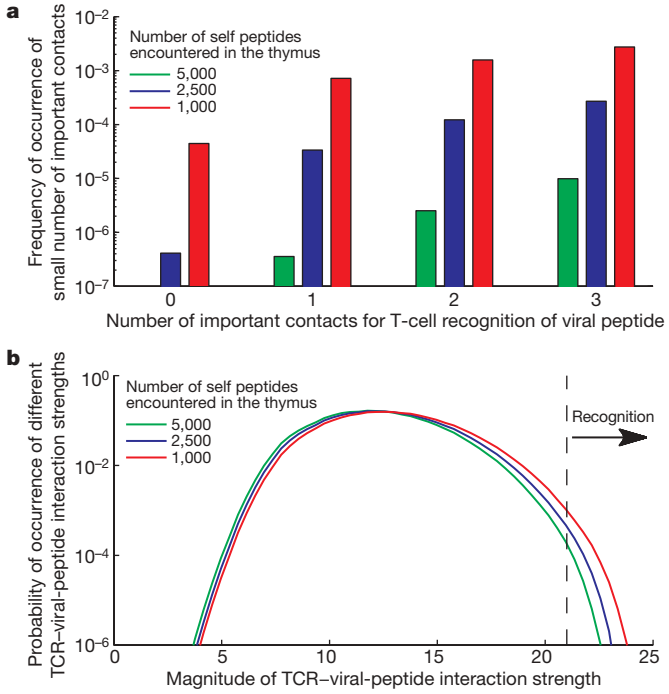

Figure 1 | Thymic selection against fewer self peptides leads to a more cross-reactive T-cell repertoire. a, Histogram of the frequency with which T cells recognize viral peptides (that is, not seen in the thymus) through only a small number (0, 1, 2, 3) of important contacts is shown for three T-cell repertoires that developed with different numbers of self-peptide–HLA complexes in the thymus. Important contacts were determined by making single point mutations. If the TCR–peptide–HLA interaction is sufficiently strong, no single point mutation can abrogate recognition, resulting in zero important contacts. A higher frequency of occurrence of a small number of important contacts indicates a more cross-reactive T-cell repertoire because only mutations at these contacts are likely to abrogate recognition. The frequency with which T cells recognize viral peptides through many significant contacts (greater than four) is larger for T-cell repertoires restricted by HLA alleles that present more self peptides in the thymus (not shown). **b**, The probability that a TCR binds to viral peptides with a certain interaction strength is shown for three T-cell repertoires (as in **a**). A particular TCR recognizes a viral peptide when the binding strength exceeds the recognition threshold (dotted black line). Members of a T-cell repertoire selected against fewer self peptides are more likely to recognize a viral peptide. The model we used describes qualitative trends robustly[9,10] (Methods), but is not meant to be quantitatively accurate.

that recognize viral peptides through many important contacts is larger for repertoires restricted by HLA alleles that present a greater diversity of self peptides in the thymus (data not shown for >four contacts). Mutations at sites different from the important contacts do not affect binding strength substantially. Therefore, when the interaction between peptide–HLA and TCR is mediated by fewer important contacts, a larger number of possible point mutations of the peptide do not affect peptide recognition, thereby making the T cells more cross-reactive to mutants that arise. Thus, the HLA-B57-restricted T-cell repertoire is expected to be more cross-reactive to mutants of targeted viral peptides than repertoires restricted by HLA alleles that present a greater diversity of self peptides.

Our computational models give this qualitative mechanistic insight, but do not provide quantitative estimates of the extent of this enhanced cross-reactivity of T cells. However, compelling experimental data[13] has shown that the effect revealed by our studies is important in humans. Peripheral blood mononuclear cells from patients expressing HLA-B57 contained CTLs that were more cross-reactive to various HIV epitopes and their point mutants than those of HLA-B8-positive patients. HLA-B8 is associated with rapid progression to disease[13], and the most accurate algorithm for peptide binding suggests that the HLA-B8 molecule binds a greater diversity of self peptides than HLA-B57 (Supplementary Fig. 4 and Supplementary Table 1). Other experimental studies also show that patients expressing HLA-B57 cross-recognize point mutants of the dominant epitope and use more public TCRs[14,15].

Next, we computed interaction strengths between diverse viral peptides and members of T-cell repertoires restricted by HLA molecules that present differing numbers of self peptides in the thymus. This allowed us to obtain the probability with which a randomly picked T-cell clone and viral peptide will interact sufficiently strongly for recognition to occur. The results (Fig. 1b) indicate that a typical $CD8^+$ T cell restricted by an HLA molecule such as HLA-B*5701, which presents fewer peptides in the thymus, has a higher probability of recognizing a viral epitope compared to a T cell restricted by other HLA molecules. Thus, more HLA-B*5701-restricted T cell clones are likely to recognize a viral epitope, making effective precursor frequencies higher in an HLA-B*5701-restricted repertoire (a strong predictor of response magnitude[16]). A greater precursor frequency for viral epitopes in the naive repertoire restricted by HLA-B57 is indicated by experimental results showing that HLA-B*5701 contributes the most to acute-phase CTL responses of all HLA alleles tested[17].

The results in Fig. 1 stem from the constraint that thymocytes must avoid being negatively selected by each self-peptide–HLA complex encountered during development in the thymus. T cells expressing TCRs with peptide contact residues composed of amino acids that interact strongly with other amino acids (for example, charged residues, flexible side chains) have a high probability of binding to a self peptide strongly. The greater the diversity of self peptides presented in the thymus, the higher the chance that a TCR with such peptide contact residues will encounter a self peptide with which strong interactions will result in negative selection. Thus, as the diversity of self peptides presented in the thymus increases, the peptide contact residues of TCRs in the mature T-cell repertoire are increasingly enriched in weakly interacting amino acids (Supplementary Fig. 5). T cells bearing TCRs with weakly interacting peptide contact residues recognize viral peptides by means of several moderate interactions, making many contacts important for recognition. In contrast, TCRs with peptide contact residues containing strongly interacting amino acids are more likely to recognize viral peptides through a few important contacts mediated by these residues, making recognition cross-reactive to mutations at other peptide sites. These mechanistic insights are supported by experimental results[7,9] (Supplementary Discussion 2).

By studying a model of host–pathogen dynamics that builds on past models of host–HIV interactions[18–20], we explored the consequences

of the HLA-B57-restricted CD8$^+$ T-cell repertoire having a higher precursor frequency for viral peptides and being more cross-reactive to point mutants of targeted epitopes on the control of HIV. Because of the importance of immune control exerted by CD8$^+$ T cells[17,21], we focused on the interaction between a mutating virus quasispecies and epitope-directed, variably cross-reactive, host CTL responses.

The essential features of the model are depicted in Fig. 2a (details in Methods). The virus is modelled as a number of epitopes consisting of strings of amino acids, and new viral strains (point mutations of epitopes), which differ in replicative fitness, arise over the course of infection. For each individual, an HLA-dependent CD8$^+$ T-cell repertoire was chosen. To mimic the results obtained from our thymic selection calculations (Fig. 1b), more or less cross-reactive repertoires were chosen (Supplementary Fig. 6) to represent HLA-B57-restricted T cells and those restricted by other HLAs, respectively. Infection rates were limited by target CD4$^+$ T cells, and CTL contraction and memory were included. Other dynamic models were studied, including one that does not incorporate target cell limitation or CTL contraction. Our qualitative results about the effects of cross-reactivity are robust to variations in parameters and model assumptions (Supplementary Figs 7–16).

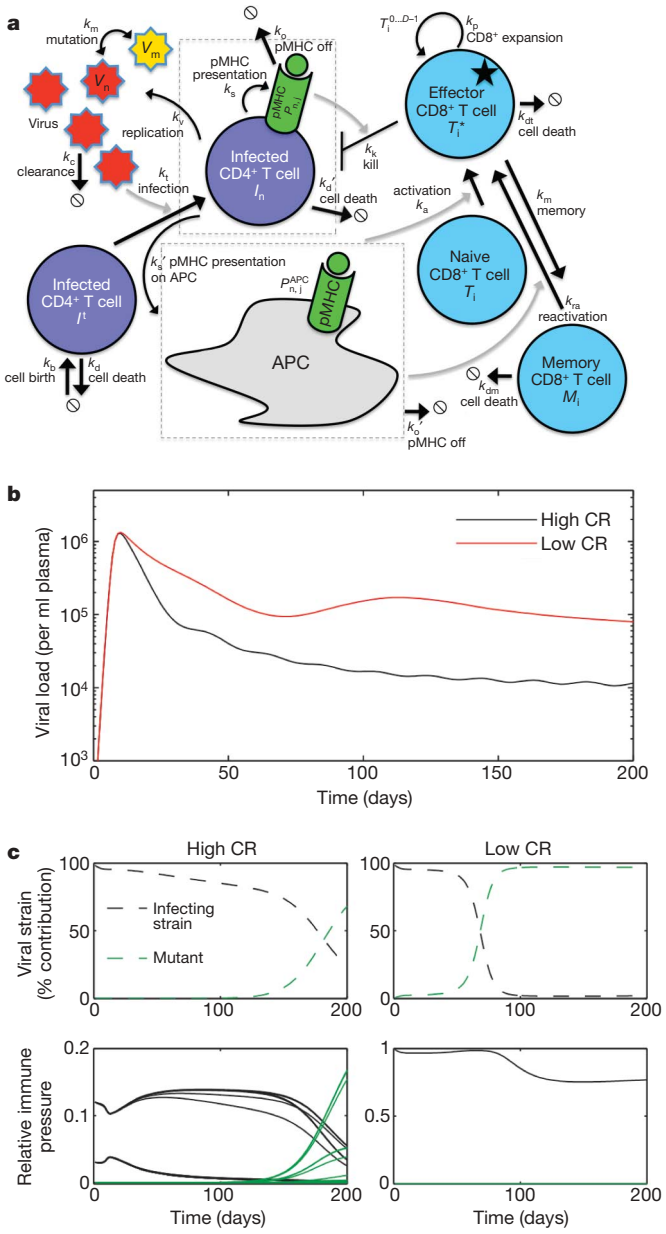

Figure 2 | Model of host–pathogen interactions shows superior viral control by cross-reactive CD8$^+$ T-cell repertoires. **a**, Dynamic model: the virus mutates, infects limited-target CD4$^+$ T cells, and is cleared. Infected CD4$^+$ T cells produce more free virus and die. Infected cells present viral peptides in complex with HLA molecules (until peptides unbind from HLA). Activated CD8$^+$ T cells produced by recognition of viral epitopes on antigen-presenting cells (APCs) proliferate and differentiate into effector CTLs. CTLs kill infected cells bearing cognate peptide–HLA complexes, and turn into memory cells that are activated after re-exposure to antigen. **b**, Simulated HIV viral loads versus time for different cross-reactivities (CR) of the CD8$^+$ T-cell repertoire. Black curve, high cross-reactivity; red curve, low cross-reactivity. Each curve is averaged over 500 simulations (each simulation represents a person). The model shows a reduced set-point viral load for people with a more cross-reactive T-cell repertoire. Other models of host–pathogen dynamics show similar effects of T-cell cross-reactivity (Supplementary Figs 7 and 8). **c**, Virus diversity and immune pressure for representative people (that is, representative simulations) with high cross-reactivity (left) and low cross-reactivity (right) of CD8$^+$ T-cell repertoires. Top panels show the relative population sizes of two dominant viral strains: the infecting strain (black), and an emerging, less fit strain (green) (other less populous viral strains are not shown). For people with a more cross-reactive T-cell repertoire, the emergent mutant strain only begins to dominate the infecting strain after 175 days, whereas for low cross-reactivity the mutant increases to nearly 100% of the viral population within 100 days after infection. Bottom panels show the relative immune pressure, defined as the rate of killing of an infected cell (see equation (4), Methods), imposed on each viral strain by different CD8$^+$ T-cell clones. Each curve represents the relative immune pressure exerted on that viral strain by a particular T-cell clone. For people with a more cross reactive T-cell repertoire, several T-cell clones exert immune pressure on both the infecting and emergent strains. For people with a low-cross-reactivity T-cell repertoire, the emergent strain is not recognized, and thus escapes.

We find that individuals with a more cross-reactive CTL repertoire control viral loads better during the acute phase of the infection (Fig. 2b). This is in agreement with findings in simian immunodeficiency virus (SIV)-infected rhesus macaques[22], where the number of cross-reactive TCR clones negatively correlates with viral load. Our simulations show that a larger number of CTL clones in a more cross-reactive T-cell repertoire recognize epitopes from the infecting viral strain (Fig. 2c). This is because the predicted higher precursor frequency for viral epitopes (Fig. 1b) leads to a greater response magnitude (as in mouse models[16]). This conclusion is supported by data showing that in people with a protective HLA allele, the initial T-cell response to HIV is dominated by T cells restricted by the protective HLA and not those restricted by other HLAs expressed[17]. Our simulations also show that enhanced cross-reactivity of the T-cell repertoire leads to greater immune pressure on the emergent viral mutants by individuals expressing HLA-B57 compared to those with T cells restricted by HLA molecules that bind more types of self peptides. The stronger immune pressure on infecting and emerging viral strains results in superior control of viral load. Thus, we predict that HIV-infected individuals with HLA alleles that bind fewer self peptides are more likely to control viral loads to low values.

To test this prediction, we studied two large HLA-typed cohorts: 1,110 controllers with less than 2,000 HIV particles ml^{-1} and 628 progressors (or non-controllers) with viral loads exceeding 10^4 ml^{-1} (Methods). From these data, we obtained the odds ratio (OR) for individual HLA alleles. People with HLA alleles associated with an OR value greater or less than one are more likely to be progressors or controllers, respectively. We focused on HLA-B alleles because they are associated with control of HIV[23]. Of 40 HLA-B alleles that were studied, significant results (P value <0.05) were obtained for five HLA-B alleles (Supplementary Table 2) and peptide-binding data are available for four of them. In support of our predictions, those HLA-B alleles associated with higher OR values also bind more self peptides (Fig. 3).

Superior control of viral load due to the greater precursor frequency and cross-reactivity of those T-cell repertoires restricted by HLA molecules that bind to few self peptides (for example, HLA-B57)

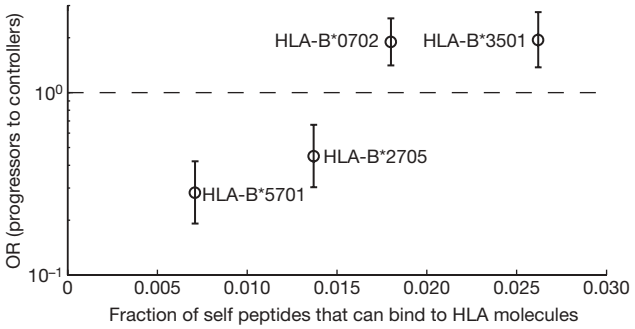

Figure 3 | HLA-B alleles associated with greater ability to control HIV correlate with smaller self-peptide binding propensities. The odds ratio (OR) for an allele is defined as: $\frac{p_w/p_{wo}}{c_w/c_{wo}}$, where p_w and p_{wo} are the numbers of individuals in the progressor cohort with and without this HLA, respectively; and c_w and c_{wo} are the numbers of individuals in the controller cohort with and without this HLA, respectively. This definition suggests that the OR measures the likelihood of an allele being correlated with progressors versus controllers, with an OR greater than one indicating association with the progressor cohort. The fraction of peptides derived from the human proteome that bind to a given HLA allele was determined using the most accurate predictive algorithms (Methods and Supplementary Table 1). Compared to experimental data, the predictive algorithms for peptide binding by HLA-B*3501 are less accurate than algorithms for the other three alleles (Supplementary Fig. 17 and Supplementary Table 1); the number reported here for HLA-B*3501 using the most accurate algorithm underestimates the binding fraction. The error bars represent the 95% confidence intervals for OR. The dotted line corresponds to equal odds for an allele being associated with progressors and controllers.

should also confer protection against diseases caused by other fast-mutating viruses. Indeed, HLA-B57 is protective against hepatitis C virus (HCV)[24], another highly mutable viral disease in which $CD8^+$ T cells are important. Also, HLA-B8, which binds a greater diversity of self peptides, is associated with faster disease progression in HCV[25] and HIV[13]. Thus, the correlation between the diversity of peptides presented in the thymus during T-cell development and control or progression of disease may be general.

Undoubtedly, many complex factors influence the relationship between HLA type and disease outcome. The effect of the new factor we have identified should be greatest for HLA molecules that bind relatively few (for example, HLA-B57) or many (for example, HLA-B7, -B35, -B8) self peptides. The strong association of HLA-B27—which binds an intermediate number of self peptides (twice as many as HLA-B57)—with viral control indicates that, in this case, the effects of T cell cross-reactivity are reinforced by this molecule binding HIV epitopes that are subject to very strong structural constraints.

Our results also point to a mechanistic explanation for as yet unexplained associations between HLA alleles that confer protection against HIV and autoimmune diseases. T cells restricted by HLA alleles that bind to few self peptides are subject to less stringent negative selection in the thymus, and should therefore be more prone to recognizing self peptides. Indeed, HLA-B57 has been associated with autoimmune psoriasis[26] and hypersensitivity reactions[27]. Enhanced cross-reactivity of HLA-B27-restricted T cells and other unique properties of this molecule (misfolding, homodimers[28]) probably contribute to the enhanced risk of autoimmunity associated with this allele[29].

Our results shed light on another intriguing observation; acutely infected patients with low viral loads (and protective HLAs) tend to target an immunodominant epitope that makes a larger relative contribution to the total CTL response as compared to individuals presenting with higher levels of viraemia[30]. This is counterintuitive as the most protective responses appear most focused, rather than broadly distributed over many epitopes. We calculated how viral load correlates with the number of CTLs responding to the immunodominant epitope divided by the total number of CTLs activated by the virus (a quantity analogous to relative contribution[30]). Mirroring experimental data, HLA alleles that restrict a more cross-reactive repertoire and are more protective also make a larger relative contribution (Supplementary Fig. 13). This result unifies the idea of both a broad and a focused response. The more cross-reactive repertoire targets more epitopes and emergent mutants, but a larger number of clones also recognize the dominant epitope (Fig. 2c).

Cross-reactive T cells are rare in people with HLA alleles that present more self peptides in the thymus than the B57 allele, but they do exist. Our results suggest that a T-cell vaccine for a diverse population must aim to activate these rare cross-reactive T cells that also target epitopes from a conserved region of the HIV genome (like HLA-B57 Gag epitopes). This will enable robust responses to infecting and mutant strains until a strain with low replicative fitness emerges, enhancing control of viral load.

METHODS SUMMARY

Predictive algorithm tools for peptide binding to HLA and Mamu molecules were obtained from the Immune Epitope Database (IEDB)[4] and were used to predict the fraction of bound peptide derived from the human and macaque proteomes[6]. Accuracies of these tools were tested on experimental data obtained from the IEDB[4]. To assess the effects of thymic selection on TCRs restricted by different MHC molecules (HLA or Mamu), we used a computational model of thymic selection described in Methods (and previously[9,10]).

To explore host–pathogen dynamics, we constructed a small model of the HIV virus with distinct epitopes and sequence diversity, based in part on past work[18–20]. We carried out numerical simulations of ordinary differential equation models, shown schematically in Fig. 2a and Supplementary Fig. 7. Parameters and their justification are given in Supplementary Tables 3 and 4 and in the Supplementary Methods. To explore cross-reactivity, we varied the distribution of pairwise-interaction free energies of TCR–pMHC contacts. Our goal was not to obtain precise numbers, but to examine the qualitative effects of variation in repertoire cross-reactivity on virus control. Qualitative results are robust to variations in parameters and assumptions (Supplementary Figs 8–16).

HLA-typed cohorts of people of diverse races were divided into HIV controllers and HIV non-controllers, and analysed for HLA association with the ability to control HIV. The results (Fig. 3 and Supplementary Table 2) were adjusted for the effects of HLA-B*0702, HLA-B*3501, HLA-B*2705 and HLA-B*5701.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 October 2009; accepted 11 March 2010.
Published online 5 May 2010.

1. Migueles, S. A. *et al.* HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl Acad. Sci. USA* **97,** 2709–2714 (2000).
2. Deeks, S. G. & Walker, B. D. Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. *Immunity* **27,** 406–416 (2007).
3. Jin, X. *et al.* Dramatic rise in plasma viremia after $CD8^+$ T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189,** 991–998 (1999).
4. Peters, B. *et al.* The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.* **3,** e91 (2005).
5. Rao, X., Fontaine Costa, A. I. C. A., van Baarle, D. & Keşmir, C. A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant $CD8^+$ T cell responses. *J. Immunol.* **182,** 1526–1532 (2009).
6. Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37,** D690–D697 (2009).
7. Huseby, E. S., Crawford, F., White, J., Marrack, P. & Kappler, J. W. Interface-disrupting amino acids establish specificity between T cell receptors and complexes of major histocompatibility complex and peptide. *Nature Immunol.* **7,** 1191–1199 (2006).
8. Huseby, E. S. *et al.* How the T cell repertoire becomes peptide and MHC specific. *Cell* **122,** 247–260 (2005).
9. Košmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M. & Chakraborty, A. K. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl Acad. Sci. USA* **105,** 16671–16676 (2008).
10. Košmrlj, A., Chakraborty, A. K., Kardar, M. & Shakhnovich, E. I. Thymic selection of T-cell receptors as an extreme value problem. *Phys. Rev. Lett.* **103,** 068103 (2009).
11. Chao, D. L., Davenport, M. P., Forrest, S. & Perelson, A. S. The effects of thymic selection on the range of T cell cross-reactivity. *Eur. J. Immunol.* **35,** 3452–3459 (2005).
12. Naeher, D. *et al.* A constant affinity threshold for T cell tolerance. *J. Exp. Med.* **204,** 2553–2559 (2007).

13. Turnbull, E. L. *et al.* HIV-1 epitope-specific $CD8^+$ T cell responses strongly associated with delayed disease progression cross-recognize epitope variants efficiently. *J. Immunol.* **176,** 6130–6146 (2006).
14. Gillespie, G. M. *et al.* Cross-reactive cytotoxic T lymphocytes against a HIV-1 p24 epitope in slow progressors with B*57. *AIDS* **16,** 961–972 (2002).
15. Yu, X. G. *et al.* Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. *J. Virol.* **81,** 1619–1631 (2007).
16. Moon, J. J. *et al.* Naive $CD4^+$ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* **27,** 203–213 (2007).
17. Altfeld, M. *et al.* HLA alleles associated with delayed progression to AIDS contribute strongly to the initial $CD8^+$ T cell response against HIV-1. *PLoS Med.* **3,** e403 (2006).
18. Althaus, C. L. & De Boer, R. J. Dynamics of immune escape during HIV/SIV infection. *PLoS Comput. Biol.* **4,** e1000103 (2008).
19. Nowak, M. A. *et al.* Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature* **375,** 606–611 (1995).
20. Wodarz, D. & Thomsen, A. R. Effect of the CTL proliferation program on virus dynamics. *Int. Immunol.* **17,** 1269–1276 (2005).
21. Cao, J. H., McNevin, J., Malhotra, U. & McElrath, M. J. Evolution of $CD8^+$ T cell immunity and viral escape following acute HIV-1 infection. *J. Immunol.* **171,** 3837–3846 (2003).
22. Price, D. A. *et al.* Public clonotype usage identifies protective Gag-specific $CD8^+$ T cell responses in SIV infection. *J. Exp. Med.* **206,** 923–936 (2009).
23. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432,** 769–775 (2004).
24. Thio, C. L. *et al.* HLA-Cw*04 and hepatitis C virus persistence. *J. Virol.* **76,** 4792–4797 (2002).
25. McKiernan, S. M. *et al.* Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* **40,** 108–114 (2004).
26. Bhalerao, J. & Bowcock, A. M. The genetics of psoriasis: a complex disorder of the skin and immune system. *Hum. Mol. Genet.* **7,** 1537–1545 (1998).
27. Chessman, D. *et al.* Human leukocyte antigen class I-restricted activation of $CD8^+$ T cells provides the immunogenetic basis of a systemic drug hypersensitivity. *Immunity* **28,** 822–832 (2008).
28. López de Castro, J. A. HLA-B27 and the pathogenesis of spondyloarthropathies. *Immunol. Lett.* **108,** 27–33 (2007).
29. Bowness, P. HLA B27 in health and disease: a double-edged sword? *Rheumatology* **41,** 857–868 (2002).
30. Streeck, H. *et al.* Human immunodeficiency virus type 1-specific $CD8^+$ T-cell responses during primary infection are major determinants of the viral set point and loss of $CD4^+$ T cells. *J. Virol.* **83,** 7641–7648 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Financial support was provided by the Mark and Lisa Schwartz Foundation, the National Institutes of Health (NIH) Director's Pioneer award (A.K.C.), Philip T and Susan M Ragon Foundation, Jane Coffin Childs Foundation (E.L.R.), the Bill and Melinda Gates Foundation, and the NIAID (B.D.W., T.M.A. and M.A.). This project has been funded in whole or in part with federal funds from the National Cancer Institute, NIH, under contract no. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Author Contributions A.K. and E.L.R. contributed equally to this work. A.K.C. and B.D.W. initiated the project. A.K., E.L.R. and A.K.C. developed the computational models. A.K., E.L.R., A.K.C. and B.D.W. analysed computational results. Y.Q., F.P., M.C., S.G.D. and B.D.W. collected and analysed the data from cohorts of HIV-infected people. A.K., E.L.R., T.M.A., M.A., M.C., B.D.W. and A.K.C. contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.K.C. (arupc@mit.edu) or B.D.W. (bwalker@partners.org).

METHODS

HLA–peptide binding predictions. There are at present several HLA–peptide binding prediction methods. The performance of these algorithms to identify new epitopes has recently been benchmarked against experimental data[31]. In general, artificial neural networks (ANN)[32] and the stabilized matrix method (SMM)[33] were found to be superior to other methods[31]. We used ANN and the SMM (versions 2009-09-01 and 2007-12-27) prediction tools provided by the IEDB[4]. Accuracy of prediction tools was tested against experimental data downloaded from the IEDB in September 2009 (Supplementary Fig. 1, Supplementary Table 1 and Supplementary Notes 1). These experimental data were obtained by two methods: competition assays, in which purified MHC and radioactive labelling are used; and association studies, in which purified MHC and fluorescence labelling are used. Data obtained from the two methods show significant correlations of measured binding affinities (as measured by half-maximum inhibitory concentration (IC_{50}) and half-maximum effective concentration (EC_{50}))[5]. Prediction tools were tested against experimental data for accuracy of classifying peptides into binders ($IC_{50} < 500$ nM) and non-binders ($IC_{50} \geq 500$ nM); the chosen thresholds are commonly accepted values[5]. We also tested how well these tools predict absolute measured affinity values, not just classification of binders and non-binders, which is dependent on the chosen thresholds. The accuracy of the prediction tools thus determined are summarized in Supplementary Table 1 and Supplementary Fig. 1. We excluded all HLA and Mamu alleles for which there was not enough experimental data (at least 50 binders and 50 non-binders) or prediction tools were not sufficiently accurate (Supplementary Notes 1). For each HLA and Mamu allele, the most accurate prediction tool was used to predict the fraction of unique peptides derived from the human and macaque proteome (Homo_sapiens.GRCh37.55.pep.all.fa and Macaca_mulatta.MMUL_1.56.pep.all.fa obtained from Ensembl[6]) that can bind to that allele. We focused only on the binding abilities of peptides of 9 amino acids to HLA molecules, because there is not enough experimental data available for the binding affinities of peptides of 8, 10 and 11 amino acids to HLA-B*5701 and the other relevant HLA-B alleles that emerged from our analyses (HLA-B*2705, HLA-B*0702 and HLA-B*3501).

Thymic selection model and antigen recognition. The TCR contact residues of peptides and the peptide contact residues of TCRs are represented as strings of sites of length N. One-million sequences of TCR peptide contact residues were subject to development in a thymus containing M self peptides with TCR contact residues generated according to their frequency of occurrence in the human proteome. A particular TCR with the sequence of peptide contact residues $\vec{t}$ successfully matures in the thymus if it avoids negative selection with all self peptides ($E_{\mathrm{int}} > E_n$) and is positively selected by at least one self peptide ($E_{\mathrm{int}} < E_p$). Interaction free energy between sequences of TCR and peptide contacts, $\vec{t}$ and $\vec{s}$, is, respectively:

$$E_{\mathrm{int}}(\vec{t},\vec{s}) = E_c + \sum_{i=1}^{N} J(t_i,s_i) \quad (1)$$

where E_c represents an interaction between a TCR and an HLA molecule, and J is an empirically determined statistical potential between interacting amino acids on a TCR and a peptide. Antigenic peptides are recognized by a mature TCR if binding is stronger than the threshold for recognition ($E_{\mathrm{int}} < E_r$). The statistical potentials do not necessarily provide quantitatively accurate values of the interaction free energies. However, theoretical analyses and computational results[9,10] show that the following qualitative result is true regardless of the choice of the statistical potentials: the smaller the diversity of self peptides presented in the thymus, the greater the cross-reactivity of the mature T-cell repertoire that develops therein. More details of the model and the insensitivity of our results to parameter variations (for example, qualitative results do not depend on the choice of J or E_c (as long as E_c is not too small or large)) are described in Supplementary Information (Supplementary Figs 2 and 3) and elsewhere[9,10]. The parameters used for the results in the main text are: $N = 5$; $E_n - E_c = -21\ k_BT$; $E_p - E_n = 2.5\ k_BT$; $E_r = E_n$ and Miyazawa–Jernigan statistical potential J[34]. Numbers of self peptides presented in the thymus, M, were varied to represent different HLA alleles.

Host–pathogen interaction dynamics. We constructed a small model of HIV with distinct epitopes and sequence diversity, based in part on models developed previously[18,19]. The virus is modelled as displaying L epitopes, each consisting of M amino acid residues that may be of N types. Different viral strains arise through point mutations at the amino-acid sites, giving $(N^M)^L$ distinct strains. The number of different pMHC types is $L \times N^M$, because peptide sequences at epitope positions $1 \ldots L$ are considered to be distinct. The system of ordinary differential equations corresponding to the model in Fig. 2 and based on previous work[20] is as follows:

$$\frac{dV_n}{dt} = k_v^n I_n - k_c V_n + k_m \sum_{n:m} (V_m - V_n) \quad (2)$$

$$\frac{dI^t}{dt} = k_b - k_d I^t - k_t I^t \sum_n V_n \quad (3)$$

$$\frac{dI_n}{dt} = k_t V_n I^t - k_d{}' I_n - \sum_i \sum_j \sigma_{i,j} k_k P_{n,j} T_i^* \quad (4)$$

$$\frac{dP_{n,j}}{dt} = k_s I_n - k_o P_{n,j} - \frac{dI_n^{(\mathrm{kill})}}{dt} \frac{P_{n,j}}{I_n} \quad (5)$$

$$\frac{dP_{n,j}^{APC}}{dt} = k_s{}' I_n - k_o{}' P_{n,j}^{\mathrm{APC}} \quad (6)$$

$$\frac{dT_i}{dt} = -k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{\mathrm{APC}} \quad (7)$$

$$\frac{dT_i^0}{dt} = -k_p T_i^0 + k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{\mathrm{APC}} + k_{ra} M_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{\mathrm{APC}} \quad (8)$$

$$\frac{dT_i^m}{dt} = 2k_p T_i^{(m-1)} - k_p T_i^m \quad (9)$$

$$\frac{dT_i^*}{dt} = 2k_p T_i^{(D-1)} - k_{dt} T_i^* - k_m T_i^* \quad (10)$$

$$\frac{dM_i}{dt} = k_m T_i^* - k_{dm} M_i - k_{ra} M_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{\mathrm{APC}} \quad (11)$$

Target $CD4^+$ T cells, I^t, are infected by free virus particles, where V_n denotes virions of strain n. I_n denotes $CD4^+$ T cells infected by virus of strain n, $P_{n,j}$ is a pMHC complex of peptide j derived from viral strain n, displayed on the surface of the infected cell, $P_{n,j}^{\mathrm{APC}}$ is a pMHC displayed by APCs and T_i is a naive $CD8^+$ T cell of clonotype i. Activated T cells undergo D rounds of cell division before becoming effector CTLs; T_i^0 is an activated $CD8^+$ T cell of type i that has not yet begun dividing and T_i^m are the dividing cells, where m runs from 1 to $D-1$. Effector CTLs, T_i^*, differentiate into memory $CD8^+$ T cells, M_i, which are activated upon re-exposure to pMHC.

If T-cell clone i recognizes pMHC j, $\sigma_{i,j}$ is 1, and 0 otherwise. In equation (2), $\sum_{n:m}$ denotes the sum over viral strains m that are Hamming distance 1 away from strain n. That is, only point mutations are allowed. The third term of equation (5) ensures that if an infected cell is killed, the pMHC bound on its surface must also disappear; $\frac{dI_n^{(\mathrm{kill})}}{dt}$ denotes the third term of equation (4), which describes killing of an infected cell by CTLs that recognize pMHC on its surface. Simulations were performed using ode45 and ode15s solvers in MATLAB. A further dynamic model, which does not incorporate target cell limitation and allows unlimited expansion of activated CTLs, was also developed to show robustness of our results to model assumptions. It is discussed in the Supplementary Information (Supplementary Figs 7–12).

Rate constants used in the models are given in Supplementary Tables 3–4, and are in keeping with values reported in the literature. We assume a concentration of 10^6 $CD4^+$ T cells per ml blood before infection, with 1% of these cells activated and thus initial targets for HIV infection[35,36]. The initial conditions of infection in the simulations were one infected $CD4^+$ T cell per ml of plasma and a naive-$CD8^+$ repertoire size of one cell per ml of each clonotype. We assume that the number of epitopes, length of each epitope, and number of amino acids (L, M, N) are all 2, giving 8 pMHC types and 16 possible viral strains. The number of $CD8^+$ clonotypes was chosen to be 20.

The interplay between antigen and immune receptor diversity is captured in this model through variability in $\sigma_{i,j}$ and viral fitness. Different fitness levels for different strains of the virus are modelled by randomly selecting k_v^n, the virus proliferation rate, for each strain from a uniform distribution between 0 and 2,000 per day[18,37], with the assumption that the infecting strain has the maximum fitness. The matrix $\sigma_{i,j}$ encodes the ability of T cells to recognize pMHCs. We generate $\sigma_{i,j}$ in such a way as to mimic the results of the thymic selection model (Fig. 1b), to investigate the effects of those predictions on host–pathogen dynamics. That is, we assume that T-cell repertoires restricted by different HLA types differ in the interaction free energies of their TCR–pMHC contacts, and generate $\sigma_{i,j}$ accordingly using a type of random-energy-like model (Supplementary Fig. 6). The interaction free energy between a T cell and an epitope is given by $\sum_a J(i,j_a)$, where $J(i,j_a)$ is

the interaction free energy between T cell of clonotype *i* and residue *a* on epitope *j*. Similar to the models used for thymic selection, the total interaction free energy is taken to be the sum of the individual residue interactions and recognition is said to occur when it exceeds a recognition threshold (in the dynamic model, T-cell sequences are not specified explicitly). $J(i,j_a)$ is a random variable chosen from a uniform distribution, and the width of the distribution determines the probability that the summed interaction energy falls above the threshold, and thus the probability that a peptide is recognized by a given T cell. Repertoires generated in this way approach a Gaussian distribution of interaction energies, and the distribution shifts and thus cross-reactivity increases when the uniform distribution from which $J(i,j_a)$ is selected is wider. Generating $\sigma_{i,j}$ in this way allows us to describe variable cross-reactivities of the T-cell repertoire (both intra- and inter-epitope), and also accounts for correlated interaction energies and thus recognition probabilities of similar peptide sequences.

HLA-allele association with ability to control HIV. SAS 9.1 (SAS Institute) was used for data management and statistical analyses. Odds ratios and 95% confidence intervals were determined using PROC LOGISTIC in a comparison of HIV controllers (those individuals who maintained viral loads of less than 2,000 copies of the virus per ml plasma on three determinations over at least a year of follow-up and, on average, for approximately 15 years[38]) to HIV non-controllers (those individuals whose viral loads exceeded 10,000 copies of the virus per ml plasma). To eliminate the confounding effects of B*0702, B*3501, B*2705 and B*5701, alleles strongly associated with progression or control, these factors were used as covariates in the logistic regression model for the analysis of all other HLA class I types[39]. All ethnic groups were included in the analyses shown (European, African-American and others) and we adjusted for ethnicity in the logistical regression model. All *P* values were corrected for multiple tests using the Bonferroni correction, a stringent and commonly used approach for multiple comparisons[40].

31. Peters, B. *et al.* A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* **2,** e65 (2006).
32. Gulukota, K., Sidney, J., Sette, A. & DeLisi, C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* **267,** 1258–1267 (1997).
33. Peters, B., Tong, W., Sidney, J., Sette, A. & Weng, Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* **19,** 1765–1772 (2003).
34. Miyazawa, S. & Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256,** 623–644 (1996).
35. Sachsenberg, N. *et al.* Turnover of $CD4^+$ and $CD8^+$ T lymphocytes in HIV-1 infection as measured by Ki-67 antigen. *J. Exp. Med.* **187,** 1295–1303 (1998).
36. Stafford, M. A. *et al.* Modeling plasma virus concentration during primary HIV infection. *J. Theor. Biol.* **203,** 285–301 (2000).
37. Parera, M., Fernandez, G., Clotet, B. & Martinez, M. A. HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions. *Mol. Biol. Evol.* **24,** 382–387 (2007).
38. Pereyra, F. *et al.* Genetic and immunologic heterogeneity among persons who control HIV infection in the absence of therapy. *J. Infect. Dis.* **197,** 563–571 (2008).
39. Hosmer, D. W., Jovanovic, B. & Lemeshow, S. Best subsets logistic-regression. *Biometrics* **45,** 1265–1270 (1989).
40. Cheverud, J. M. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87,** 52–58 (2001).

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:

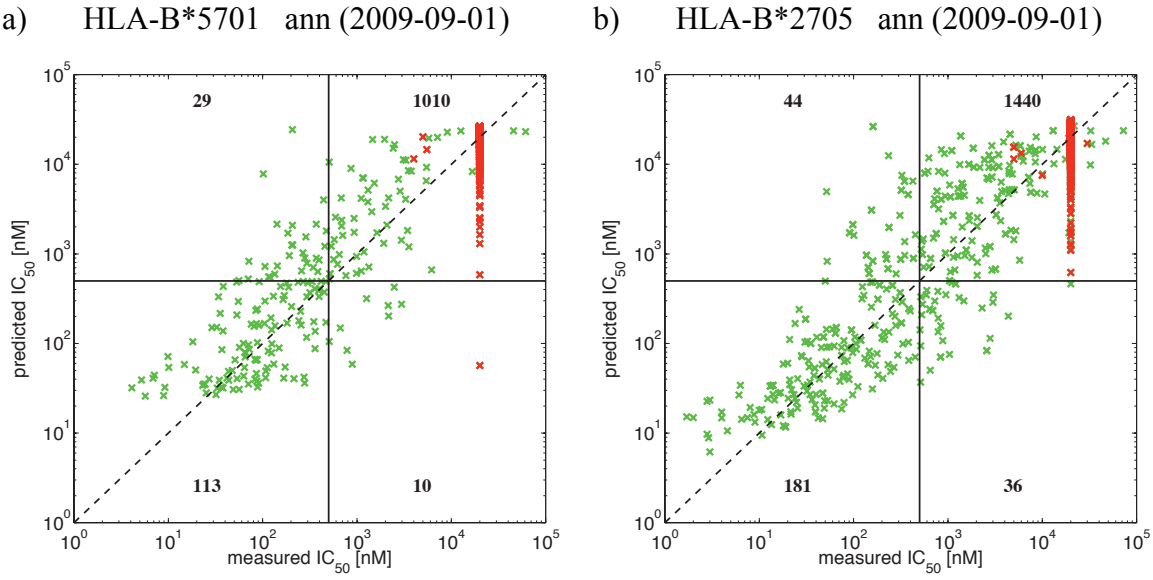

Figure Legend S1: Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*5701 allele (a) and HLA-B*2705 (b). For both alleles the best predictive algorithm (Table S1) was used. Green data points correspond to measurements, which report 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:

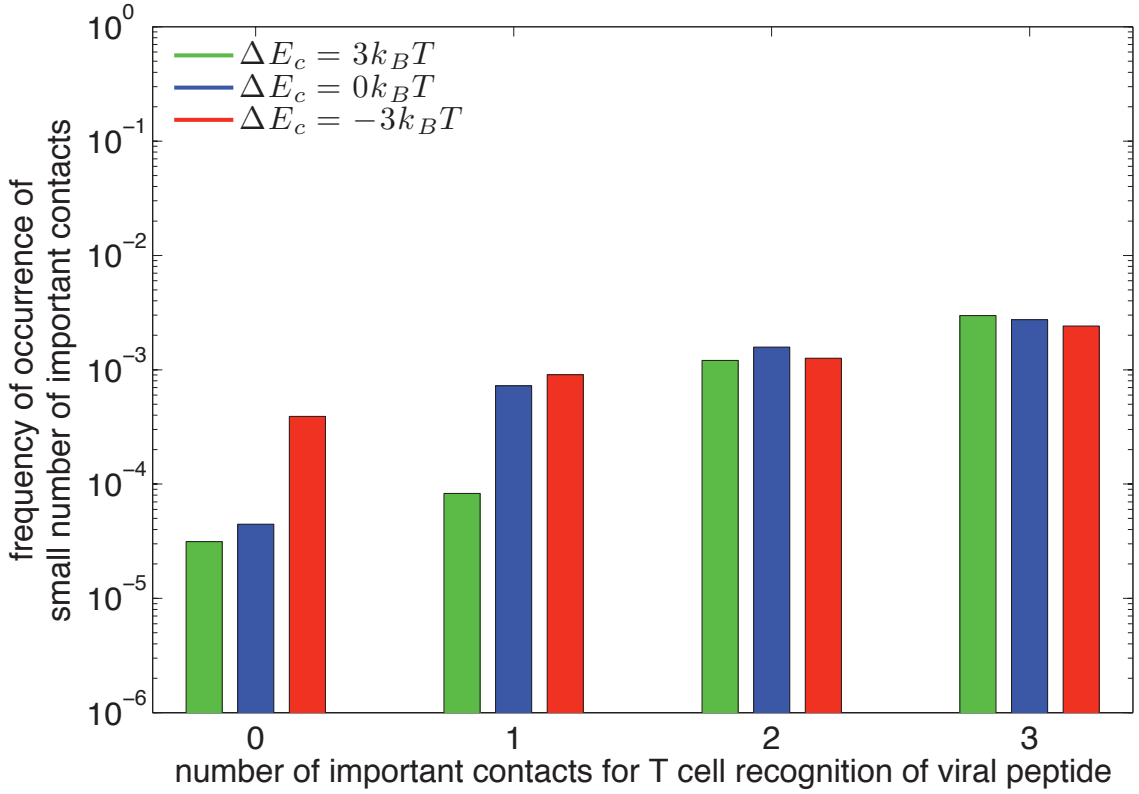

Figure Legend S2: Distribution of the number of important contacts for TCR recognition of antigenic peptides is invariant to variations in interaction free energy (E_c) between TCRs and MHC as long as these interactions are not too strong or too weak. As shown previously[9,10] too strong or weak TCR-MHC 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:

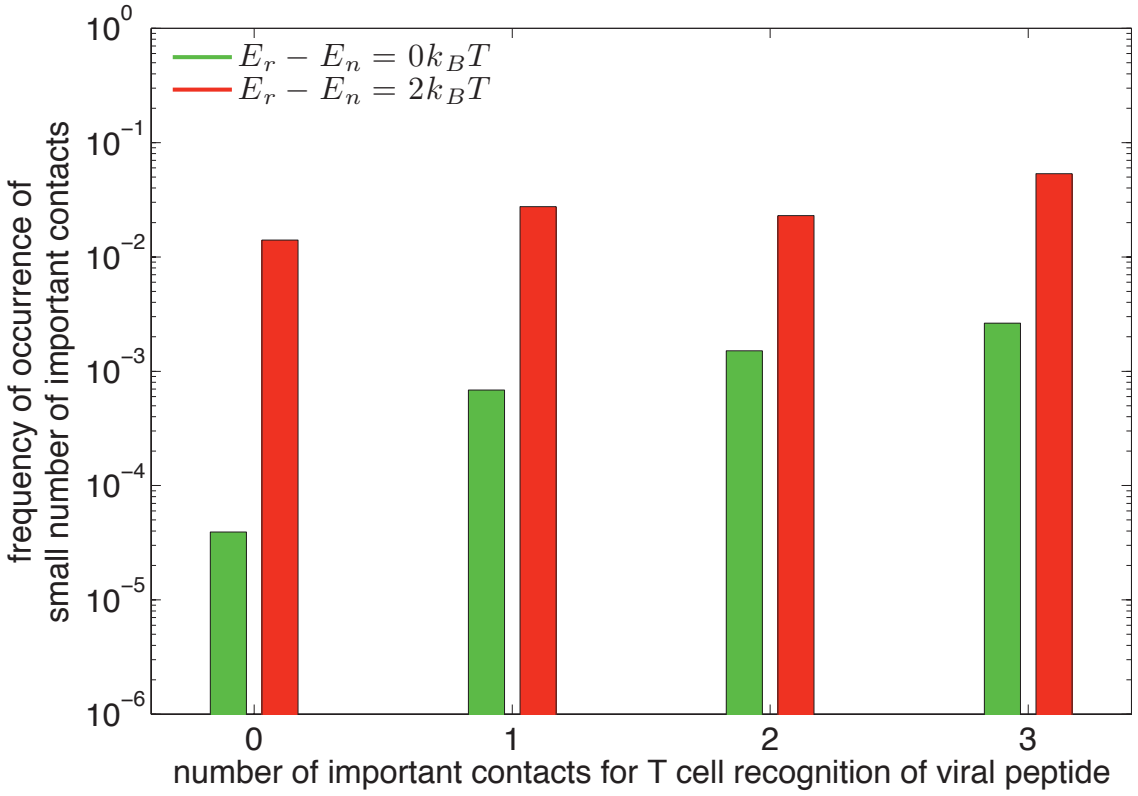

Figure Legend S3: Weaker binding free energy threshold for antigen recognition than the negative selection threshold in the thymus results in more cross-reactive TCRs. Histogram of important contacts for TCR recognition of antigenic peptides for different binding thresholds of antigen recognition for 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)

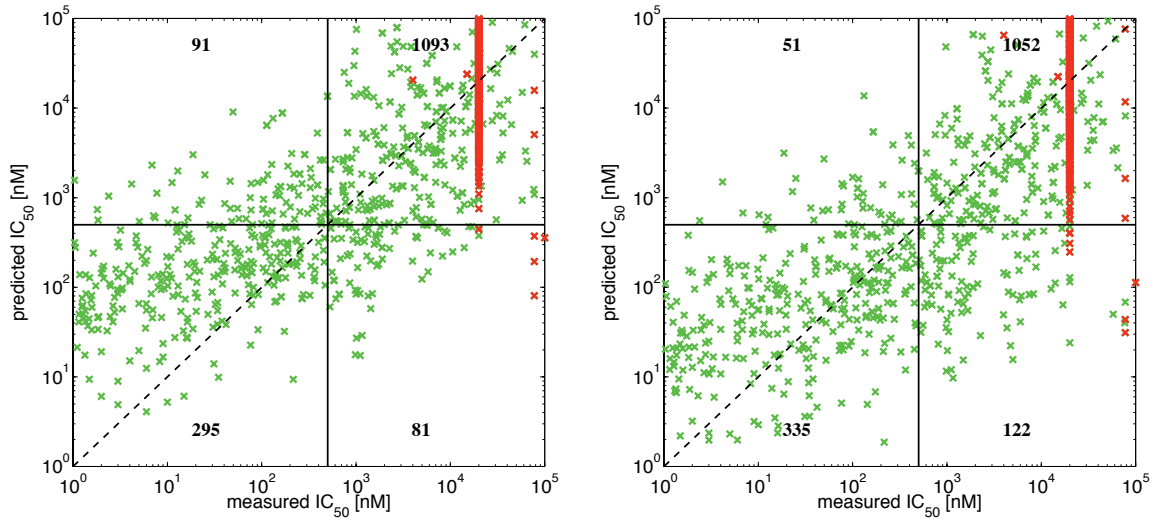

Figure Legend S4: The predictive algorithms for HLA-B*0801 are not very accurate (see also Table S1). Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*0801 allele for two predictive algorithms: smm (2009-09-01) on left and smm (2007-12-27) 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:

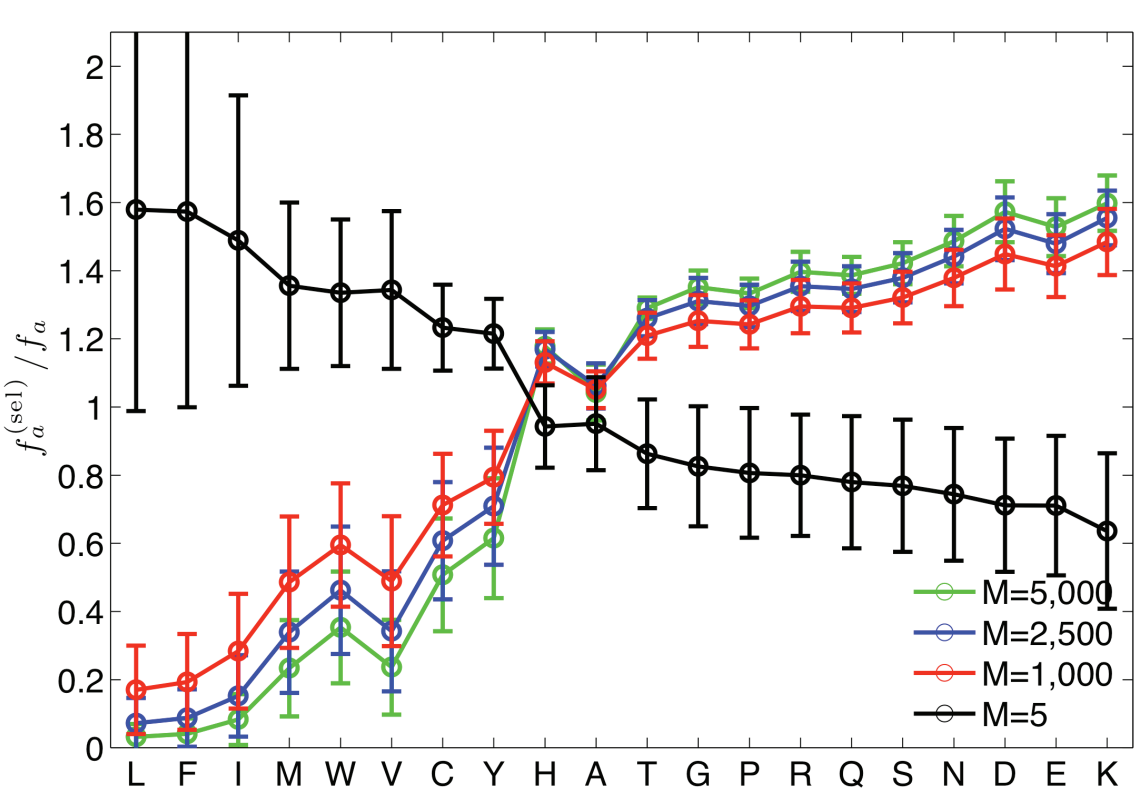

Figure Legend S5: Selection against a greater diversity of peptides (*M*) in the thymus results in selected TCRs with peptide contact residues that are more enriched in amino acids that interact weakly with other amino acids. The ordinate is the ratio of the frequencies of occurrence of an amino acid 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:

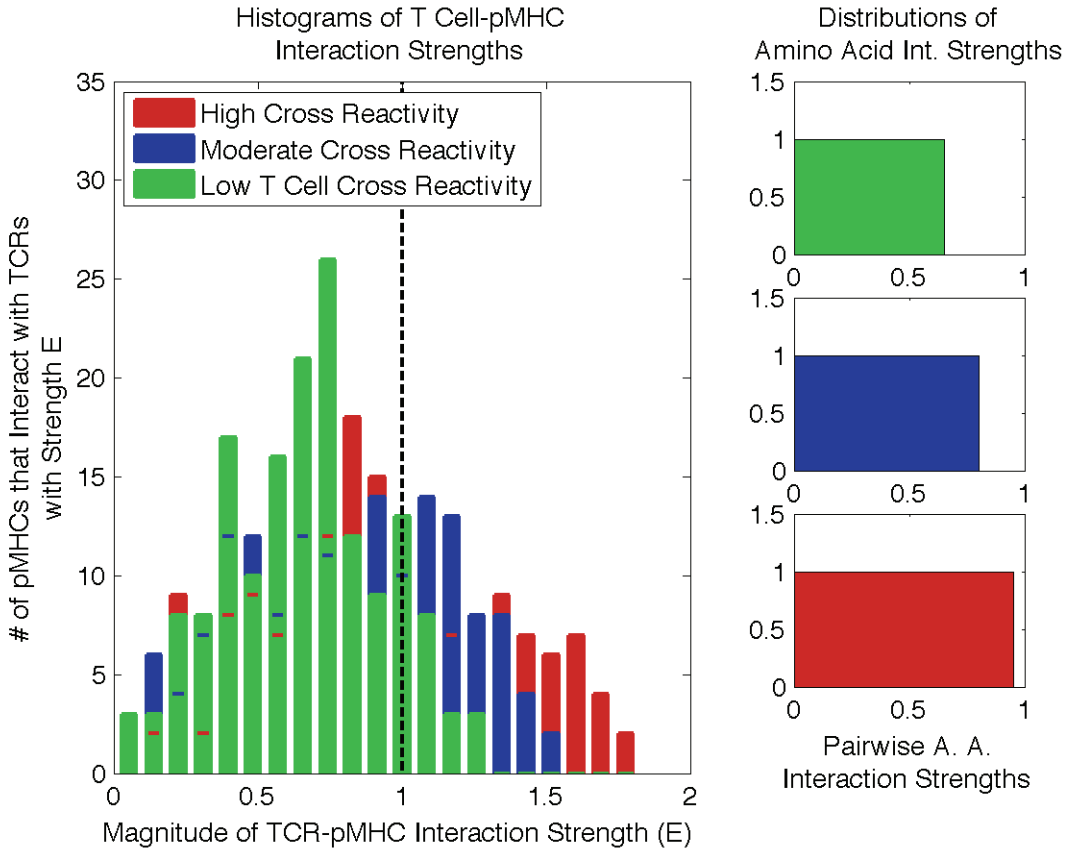

Figure Legend S6: Random energy-like model for generating $\sigma_{i,j}$, the matrix describing recognition of pMHCs by $CD8^+$ T cells. The degree of cross-reactivity in the simulation depends on the uniform distribution from which interaction strengths between individual epitope residues and the TCRs 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:

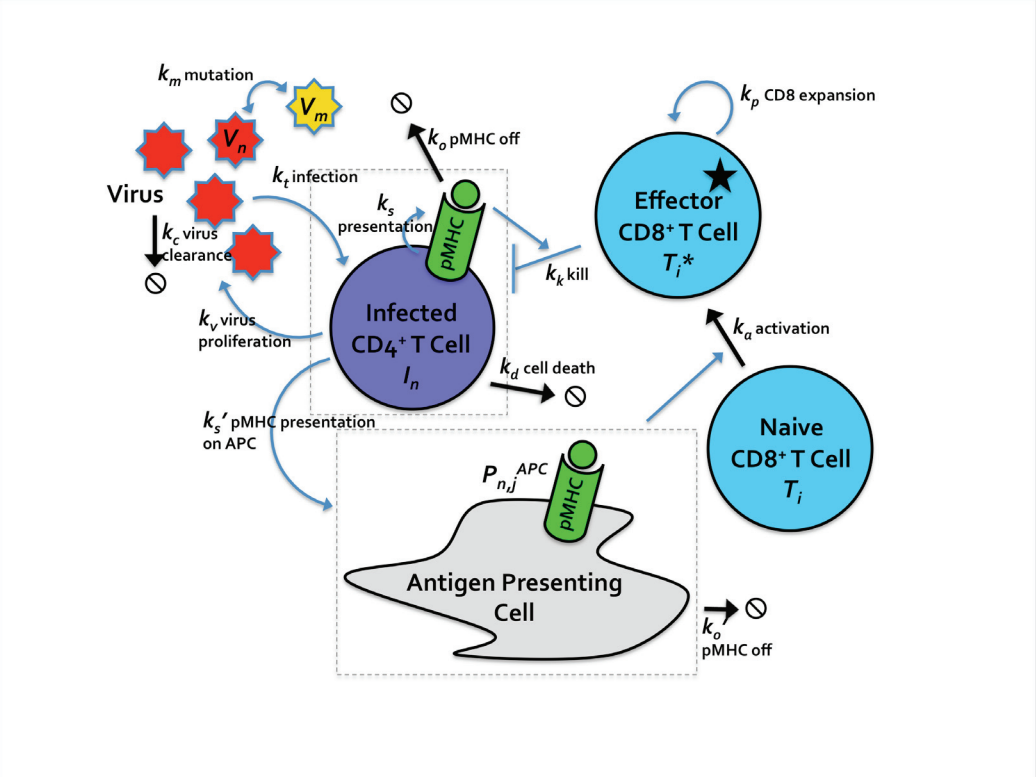

Figure Legend S7: Schematic of a model for host-pathogen dynamics that is simpler than that shown in Fig. 2 (henceforth termed, simplified model). The virus mutates, infects target CD4⁺ T cells, and is cleared. Infected CD4⁺ T cells produce more free virus, and die. Infected cells present viral peptides 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:

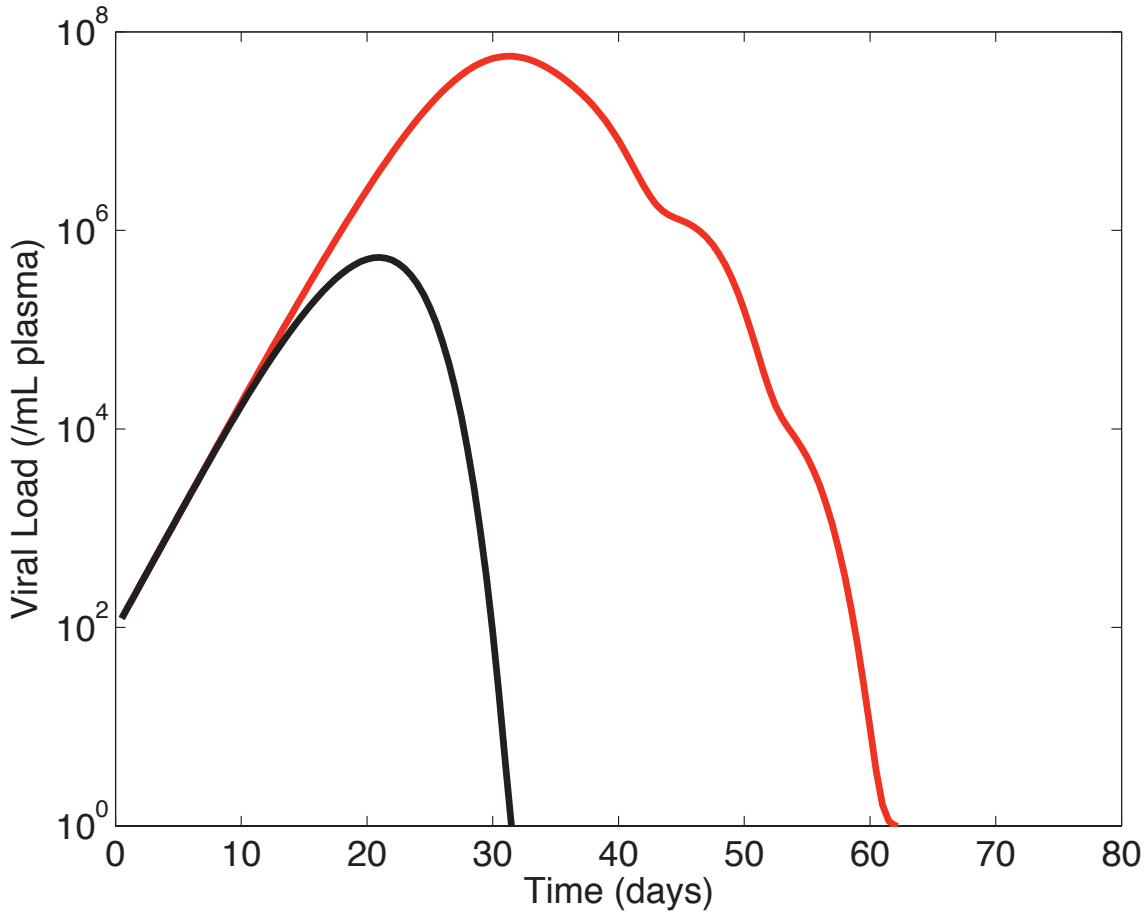

Figure Legend S8: Simulation results using the simplified model. HIV viral loads versus time for different cross-reactivities (CR) of the $CD8^+$ T cell repertoire, corresponding to the model in Fig. S7. Black curve: highly cross-reactive case. Red curve: lower cross-reactivity. Each curve is averaged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

Supplementary Figure S9:

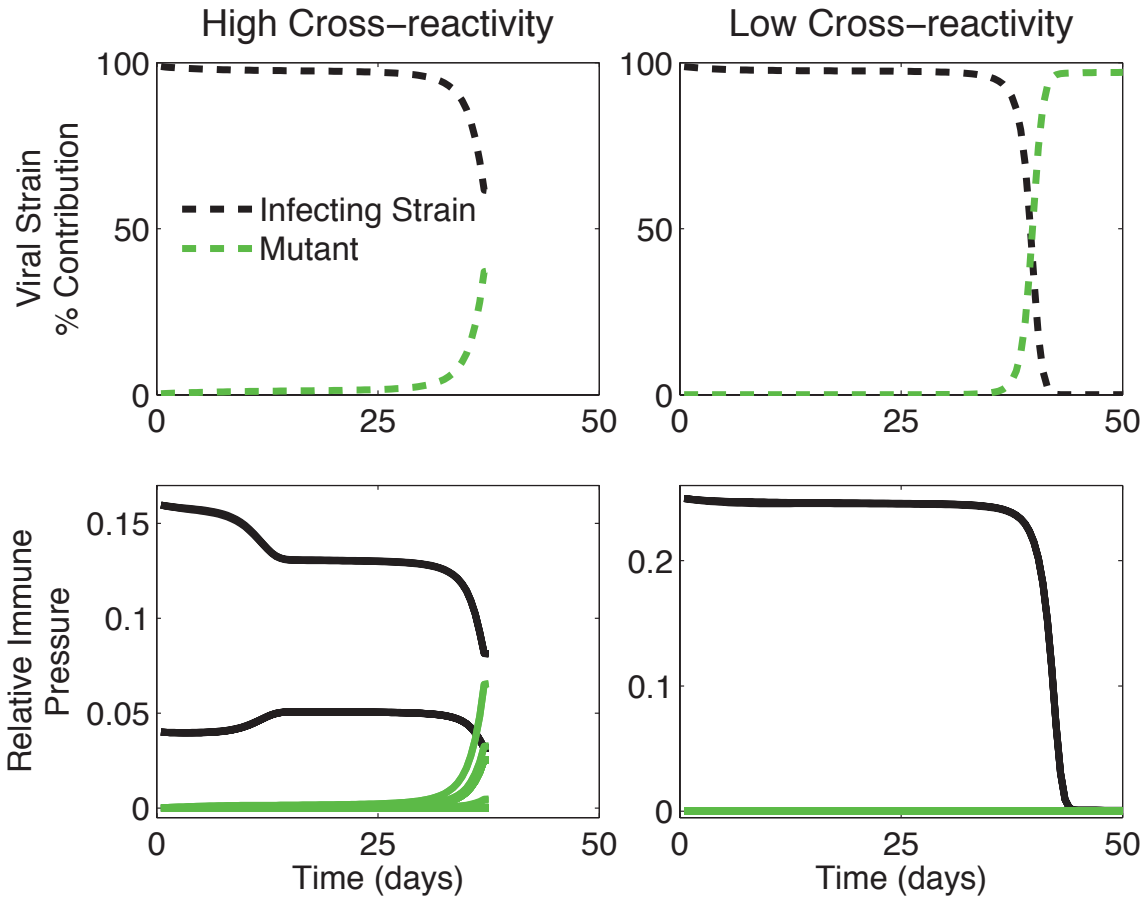

Figure Legend S9: As in Fig. 2c, but for the simplified model (schematic in Fig S7). When more clones recognize the infecting and emerging strains (left, bottom), the emerging mutant strain (green) is kept in check (left, top). However, when cross-reactivity is low, the likelihood that the mutant strain 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:

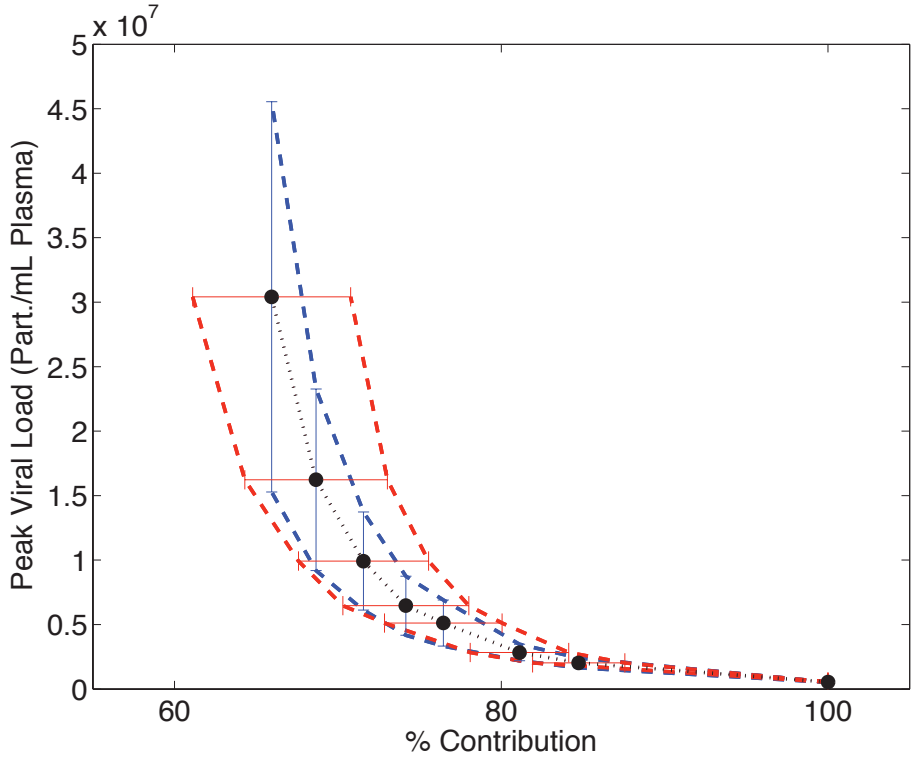

Figure Legend S10: Anticorrelation of simulated peak viral loads with percent contribution of the dominant epitope to the total CTL response for the model in Fig. S7. Percent contribution is calculated as the number of activated CTLs recognizing the immunodominant epitope over the total number of activated 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:

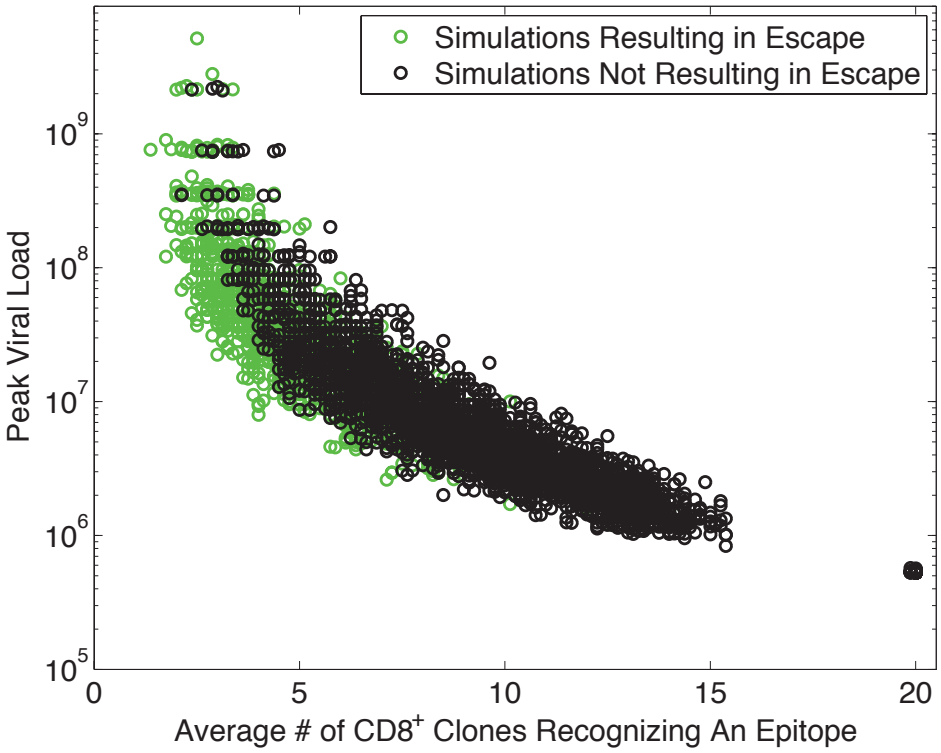

Figure Legend S11: Peak viral load versus average number of $CD8^+$ clones recognizing a pMHC in each simulation in the simplified model (schematic in Fig. S7). "Escape" is taken to mean that the population size of a mutant viral strain has become larger than that of the infecting strain at some point 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:

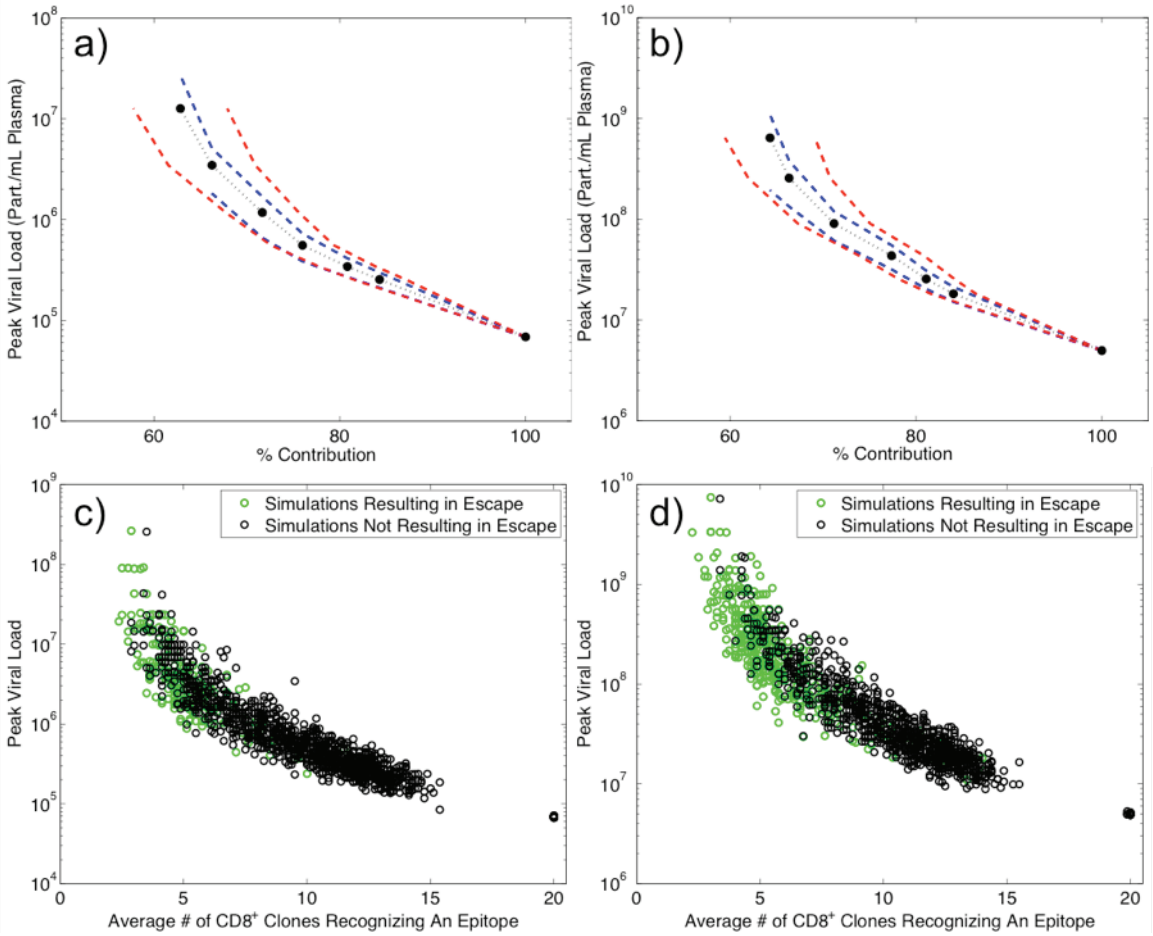

Figure Legend S12: Insensitivity of qualitative results to changes in $CD8^+$ T cell activation rate for the simplified model (schematic in Fig. S7). Left panels show simulation results with varying cross reactivity for activation rate $10 * k_a$ (k_a given in Table S4), while right panels show results 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:

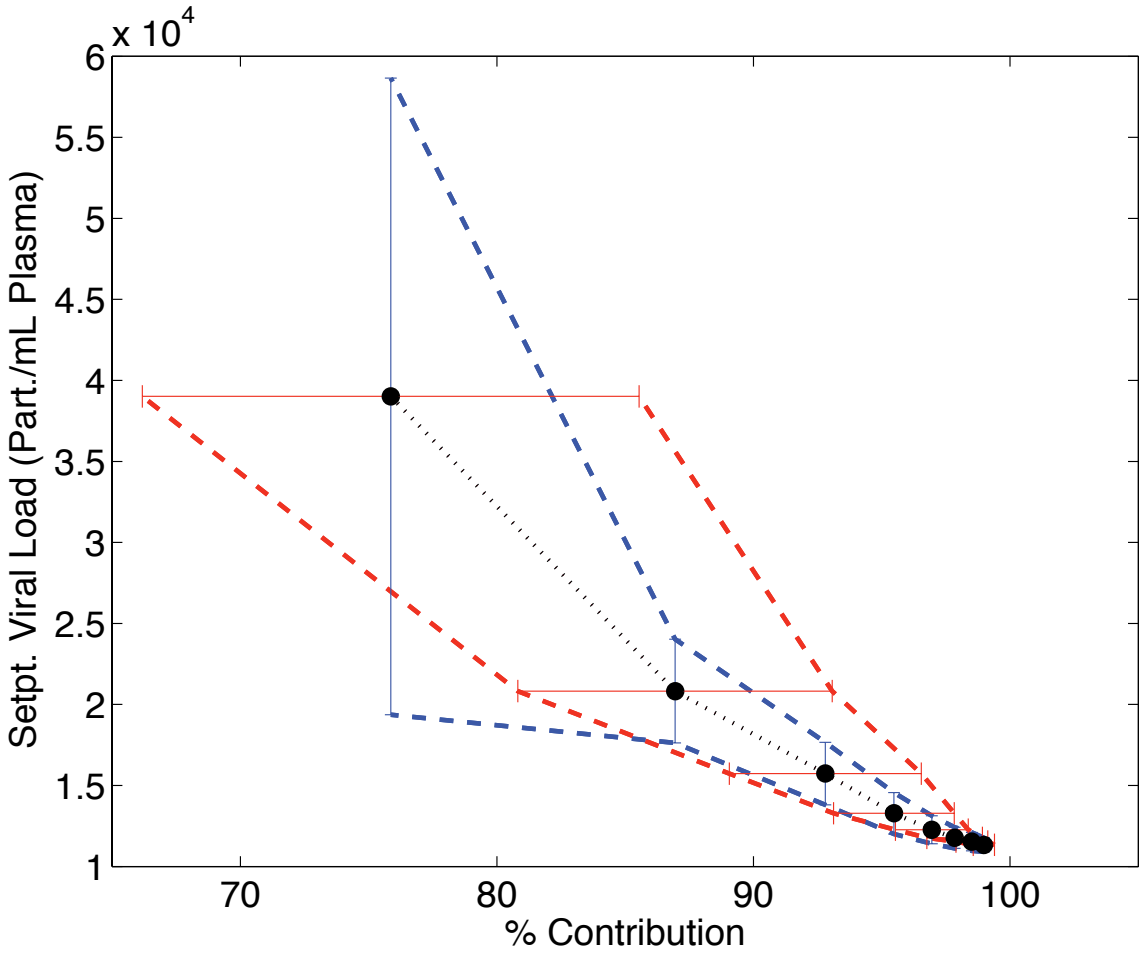

Figure Legend S13: Anticorrelation of simulated viral loads with percent contribution of the dominant epitope to the total CTL response, as in Fig. S10, but corresponding to the model discussed in the main text (Fig. 2a). Viral load and % contribution were calculated at day 200 in the simulations, to 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:

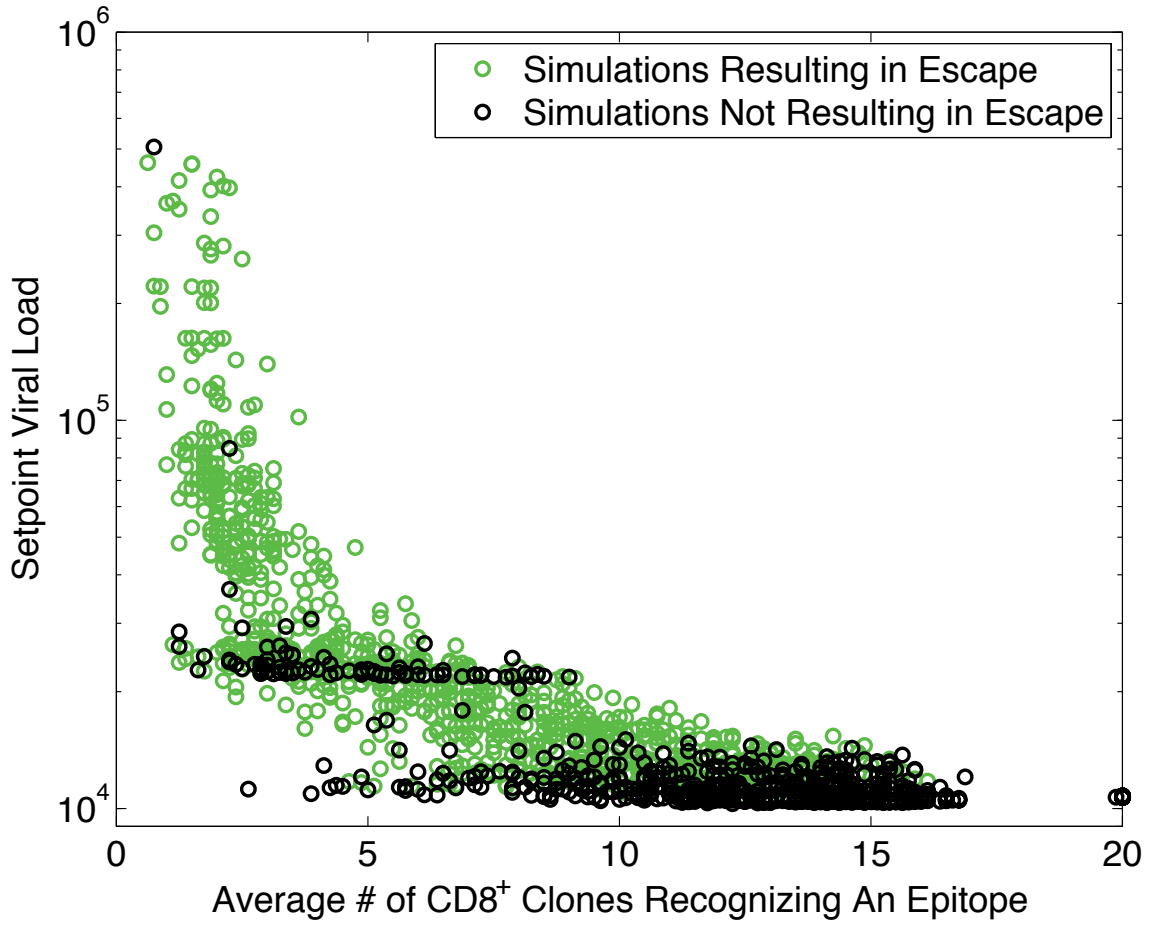

Figure Legend S14: As in Fig. S11, but for the model described in the main text (schematic in Fig. 2a). As the number of clones recognizing a pMHC increases, the setpoint viral load and probability of escape decrease.

Supplementary Figure S15:

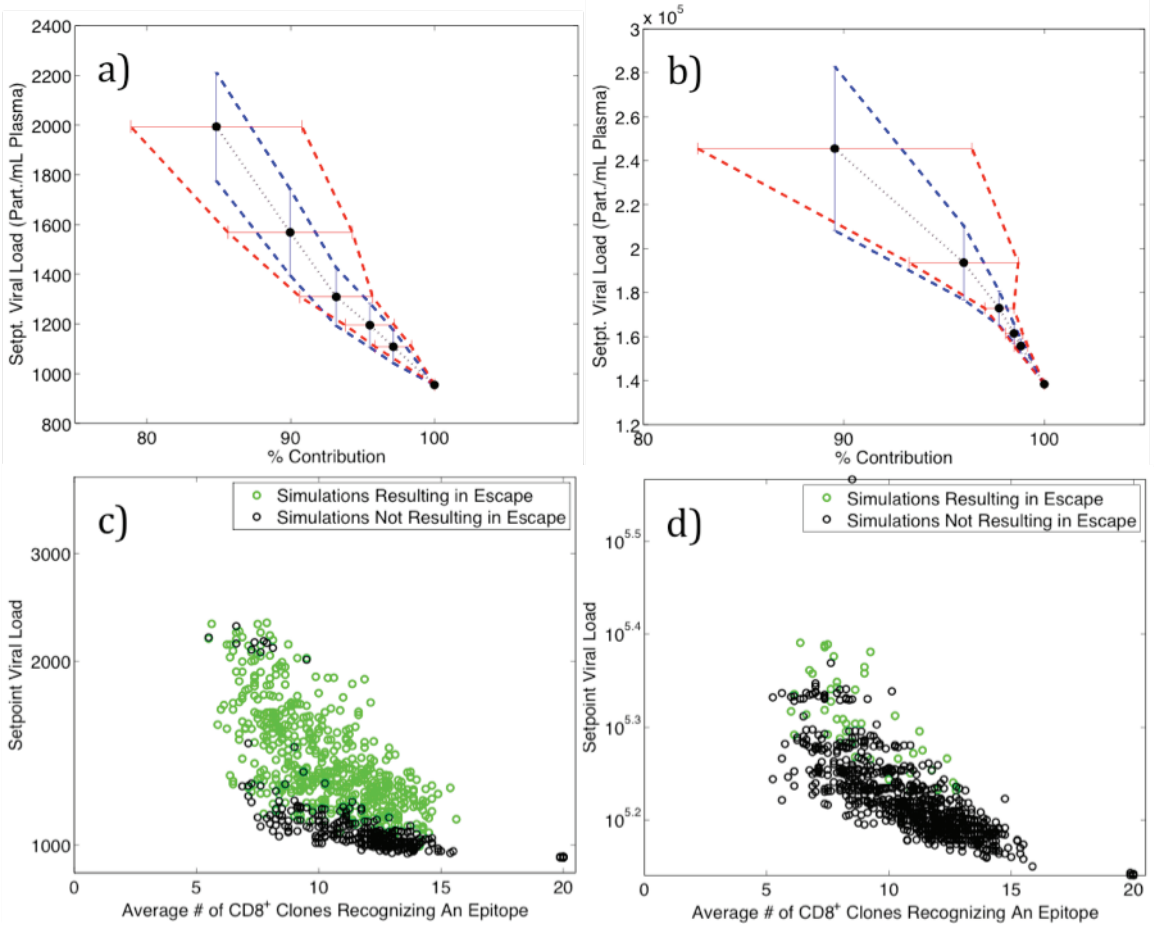

Figure Legend S15: Insensitivity of qualitative results to changes in $CD8^+$ T cell activation rate for the model described in the main text (schematic in Fig. 2a). Left panels show simulation results with varying cross reactivity for activation rate $10 * k_a$ (k_a given in Table S3), while right panels 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:

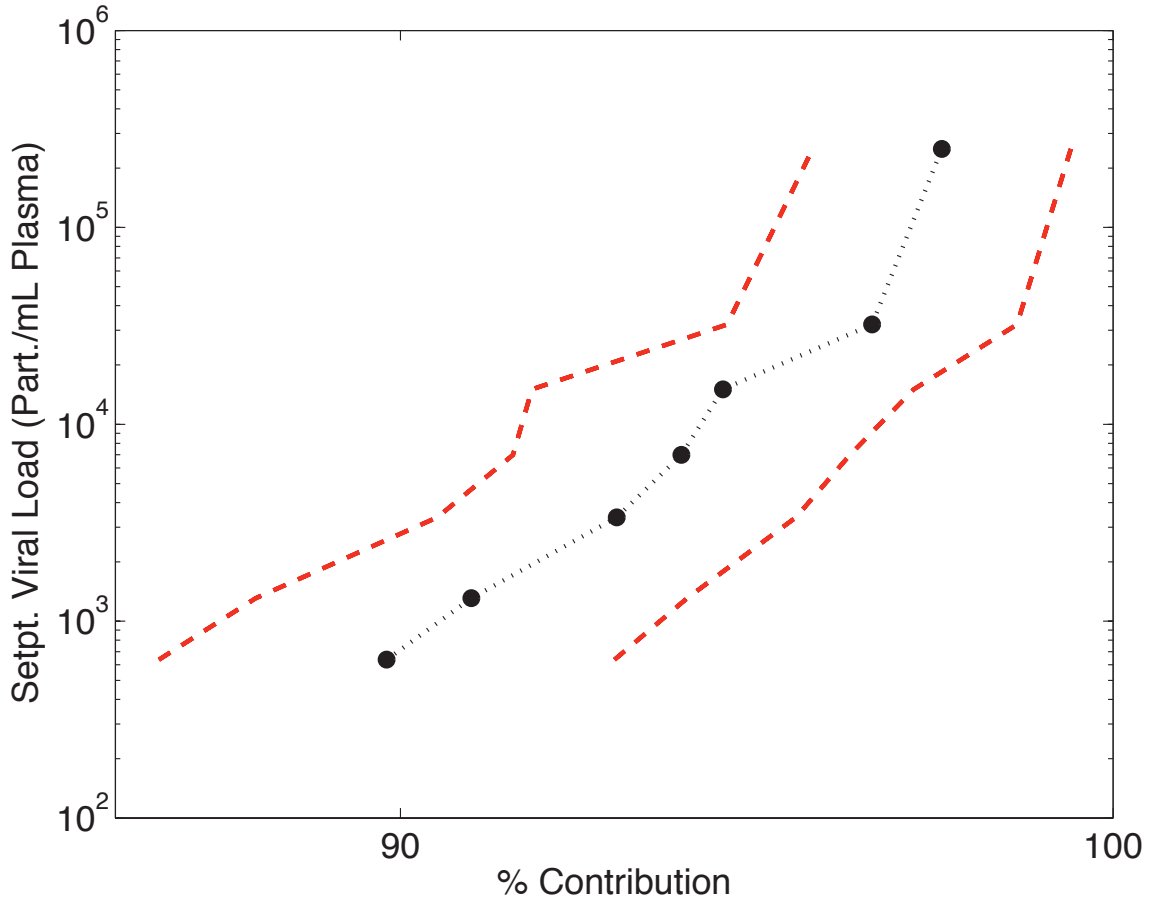

Figure Legend S16: Setpoint viral load versus % contribution of immunodominant epitope, as in Figs. S10 and S13, but where the rate of peptide presentation k_s (not cross-reactivity), is varied. Points correspond to k_s values of 200, 100, 40, 20, 10, 5, and 1 (day^{-1}), with faster presentation of 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:

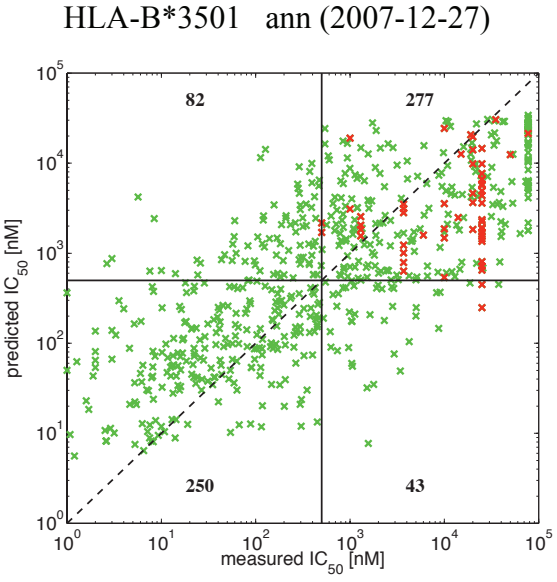

Figure Legend S17: The predictive algorithms for HLA-B*3501 are less accurate than that fo HLA-B*5701, HLA-B*0702, and HLA-B*2705 (see also Table S1). Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*3501 allele for the most accurate 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).