Supporting Information

TCR contact residue hydrophobicity is a hallmark of immunogenic CD8$^+$ T cell epitopes

Diego Chowell[a,b,1], Sri Krishna[b,d,1], Pablo D. Becker[e], Clément Cocita[e], Jack Shu[f], Xuefang Tan[f], Philip D. Greenberg[f], Linda S. Klavinskis[e,2], Joseph N. Blattman[c,2], and Karen S. Anderson[b,2]

[a]Simon A. Levin Mathematical, Computational and Modeling Sciences Center, [b]Center for Personalized Diagnostics, [c]Center for Infectious Diseases and Vaccinology, and [d]School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ 85004; [e]Department of Immunobiology, King's College London, London SE1 9RT, United Kingdom; and [f]Department of Immunology, University of Washington, Seattle, WA 98195.

[1]These authors contributed equally to this work.

[2]To whom correspondence should be addressed. Email: Karen.Anderson.1@asu.edu, Joseph.Blattman@asu.edu, or linda.klavinskis@kcl.ac.uk.

Contents:

Supporting Materials and Methods

Figs. S1-S5

Tables S2-S7

Supporting Materials and Methods

Construction of datasets. All MHC class I peptides used in this study for analyses and the design of ANN-Hydro prediction model were retrieved from IEDB (1) (www.iedb.org, last accessed: 08/11/2014). The IEDB is the largest curated dataset of MHC-I peptides identified from different primary research studies from over 334 different source organisms. We set the "Immune recognition context" as T cell response and selected "MHC class I" as the criteria for data retrieval. In total, there were 28,444 T cell epitopes reported to be immunogenic by T cell assays, including self and pathogenic epitopes and 6,142 peptides were reported to be positive by ligand elution analysis (either mass spectrometry or HPLC). To avoid redundancy and overrepresentation bias, we excluded all duplicate peptides, so that each peptide is present only once in the dataset. Positive CTL epitopes represent the immunogenic epitope group. Ligand eluted MHC-I self-peptides are generally eluted from cell surface and therefore they have been antigenically processed and MHC-bound. A vast majority of eluted self-peptides are derived from endogenous proteins. To completely separate immunogenic and non-immunogenic datasets, any immunogenic eluted self-peptide associated with autoimmunity or cancer was excluded. The remaining peptides were used as the non-immunogenic peptide dataset for our analyses. Additionally, we removed any pathogen derived non-self- eluted peptides from the eluted peptide dataset to generate mutually exclusive datasets. These unique peptides were further annotated for antigen name, peptide starting position, peptide ending position, and MHC restriction, which were required for inclusion. Peptides with "undetermined class I alleles" were also excluded. These filtering criteria resulted in a final dataset of 5,035 8-11mer immunogenic epitopes and 4,853 8-11mer non-immunogenic peptides (Table S1).

Amino acid frequency analysis. Overrepresentation of certain amino acids in immunogenic peptides was identified by calculating probability ratios of amino acids given by P(x I immunogenic)/P(x I non-immunogenic), where P(x I immunogenic) and P(x I non-immunogenic) correspond to probability mass functions and x is an amino acid. Individual amino acid probability mass functions were calculated from their frequency distributions of immunogenic epitopes and non-immunogenic peptides. Spearman's rank

correlations were quantified between probability ratios and biochemical properties (hydrophobicity, polarity, or bulkiness) of amino acids using the described amino acid scales.

Position-based hydrophobicity analysis. We transformed our datasets of immunogenic and non-immunogenic peptides into numeric arrays using the R statistical software (2). Separate numeric arrays were generated for immunogenic and non-immunogenic 8, 9 and 10mers. Mean hydrophobicity of immunogenic and non-immunogenic peptides at each position was calculated and were compared residue-by-residue through Wilcoxon rank-sum tests to quantify statistical significance.

Rate analysis of predicted peptides. An efficient prediction algorithm identifies consistently all possible CTL epitopes from a given protein in the fewest number of "hits" consistently. For each test protein, we created a subset with unique CTL epitopes retrieved either from the IEDB database. Each predicted peptide starting from rank one was queried for an exact match in the dataset of CTL epitopes. When there was an exact match, a positive hit was recorded. Graphical representations comparing the rate of predictions by the IEDB-consensus binding prediction algorithm and hydrophobicity-based predictions were generated (Fig 3).

Hydrophobicity-based ANN prediction model (ANN-Hydro). The R neuralnet package was used to design and train the two ANN-Hydro models on H-2D^b and HLA-A2 restricted 9mer peptides known to be immunogenic (n=204 and n=374, respectively) or non-immunogenic (n=232 and n=201, respectively)**.** Each peptide sequence in the respective H-2D^b and HLA-A2 datasets were transformed into a corresponding numeric sequence based on the hydrophobicity value of amino acids. Training peptides were derived from IEDB and SYFPEITHI's epitope database (1, 3). A three-layer fully connected feed-forward ANN was comprised by nine input neurons, one hidden layer with three neurons, and one output variable (Fig. S3).

Our ANN-Hydro prediction model is given by the following mathematical framework:

$$y(H) = f\left(w_0 + \sum_{i=1}^{3} w_i \cdot f(w_{0i} + W_i^T H) \right),$$

where w_0 denotes the intercept of the output neuron and w_{0i} the intercept of the i^{th} hidden neuron. Additionally, w_i denotes the synaptic weight corresponding to the synapse starting at the i^{th} hidden neuron and leading to the output neuron. $W_i = (w_{1i}, w_{2i}, ..., w_{9i})$ is the vector of all synaptic weights corresponding to the synapse leading to the i^{th} hidden neuron, and $H = (h(R_1), h(R_2), ..., h(R_9))$ the vector of all inputs, which corresponds to the numeric hydrophobicity representation of a 9mer peptide, where $h(R_i)$ is the hydrophobicity value of the amino acid R_i. Finally, the output variable $y(H)$ denotes the probability of a peptide being immunogenic (p-ANN-Hydro). Since the starting values for the weights are drawn from the standard normal distribution, the outputs were averaged over 60 realizations. The activation function $f(v)$ was chosen to be the sigmoid function $f(v) = 1/(1+e^{-v})$, and the sum of squared errors was used for the error function. The learning procedure was the resilient back-propagation with learning rate set to 0.01; a threshold set to 0.01 was defined for the partial derivatives of the error function.

Application of ANN-Hydro. For each H-2D^b and HLA-A2 restricted epitope prediction, we used the MHC-binding prediction tool IEDB-consensus to generate a list of epitope predictions on which the immunogenicity model could be applied. We normalized prediction binding scores (percentile rank) using the expression $S_{B_i} = (\delta_i - \delta_{\min})/(\delta_{\max} - \delta_{\min})$ where S_{B_i} represents the normalized score of a given peptide; δ_i, the assigned output score by IEDB-consensus; $\delta_{\min}$, the minimum score assigned in prediction output by IEDB; and $\delta_{\max}$, the maximum score assigned in the entire prediction output by IEDB. To remove poor binding peptides from the list, a subset of predicted peptides was selected by defining a S_B -threshold of 0.2 for antigen length <= 100 aa's and a S_B -threshold of 0.1 (10th percentile of predicted binders) for antigen length >100 aa's. Independently, probabilities of immunogenicity were obtained by applying the ANN-Hydro model to this subset of binding predictions. Normalized scores (S_I) were then

assigned based on these probabilities of immunogenicity. Within the spectrum of predicted binders, we prioritized epitope re-ranking based on both S_B and S_I scores with first priority given to high-immunogenicity high-binders (probability of immunogenicity >= 0.4 and S_B <=0.05; region I in Fig. S3), followed by modest-immunogenicity high-binders (probability of immunogenicity < 0.4 and S_B <=0.05; region II in Fig. S3), then high-immunogenicity modest-binders (probability of immunogenicity >= 0.4 and S_B > 0.05; region III in Fig. S3), and modest-immunogenicity modest-binders (probability of immunogenicity < 0.4 and S_B > 0.05; region IV in Fig. S3). For the antigens with length <= 100 aa's, the S_B cutoff for the four regions was set to 0.1 and probability of immunogenicity threshold remained at 0.4. Predicted peptides in each section were re-ranked based on a total score defined as $S = S_B \cdot S_I$. Final ranked list was obtained by sequential appending of the re-ranked peptides from each region. The list of predicted peptides was ranked based on this total score ranging from lowest score to the highest score. The lower the total score of a predicted peptide, the higher its probability of being an immunogenic epitope. Workflow of the prediction strategy is shown in Fig. S3.

Statistical analysis of predicted CTL epitopes. We used the F-test to quantify statistical significance (P < 0.05) of the variation of predicted rankings of T cell epitopes across different antigens between ANN-Hydro together with IEDB-consensus and IEDB-consensus alone.

***In vivo* discovery of HIV-1 Gag epitopes**

Mice. C57BL/6 mice were obtained from Harlan Laboratories. All mice used were between 6 and 8 weeks of age. All animal study protocols were conducted in accordance with guidelines approved by the Institutional Animal Care and Use Committee at Kings College London and in full compliance with UK Home Office regulations under a project license to L.S.K.

Vaccine immunization. Codon optimized HIV-1 gag plasmid DNA ZM96 from strain 96ZM651.8 (provided by B Hahn, through the Centre for AIDS Reagents [CFAR] UK) and codon optimized HIV-1 gag

Consensus B plasmid DNA (provided by D Garber, Emory University, USA) were used to construct and propagate replication defective (E1, E3 deleted) recombinant Adenovirus type 5 (rAdHu5) vectors as described previously for the HIV-1 gag strain 97CN54(4). Animals were immunized with 10^9 virus particles (vp) as determined by the DNA Pico-Green assay (Invitrogen) and administered either i.m. in the quadricep muscle (rAdHu5 Consensus B gag) or i.d. at the base of the tail (rAdHu5 ZM96 and rAdHu5 CN54).

Peptides. 15mer peptides with an 11 amino acid overlap spanning the HIV-1 CN54 Gag protein and a 20mer set of peptides with 10 amino acid overlap spanning HIV-1 ZM96 were provided by CFAR, a set of 15mers with an 11 amino acid overlap spanning the HIV-1 Consensus subtype B Gag protein were provided from the NIH AIDS Reagent and Reference Program. 'Optimal' 9mer or 11mer peptides from HIV-1 CN54 Gag, ZM96 Gag and HIV-1 Consensus B were purchased from Proimmune.

T cell epitope mapping by intracellular interferon gamma staining. Spleens were harvested 14 days after immunization, homogenized to single-cell suspensions, and RBCs were lysed using ACK lysis buffer (Lonza). Splenocytes were then used for *in vitro* re-stimulation, where 10^6 cells were incubated for 6 h at 37°C with anti-CD28 (2μg/ml; BD Pharmingen), either alone (unstimulated control) or with peptides, either in pools or individually (each at 1μg peptide/ml), derived from Consensus B Gag. Brefeldin A (10μg/ml, Sigma-Aldrich) was added for the last 5 h of culture. After washing, cells were stained with anti-CD8 (clone 37.51, BD Biosciences) for 20 min, then fixed and permeabilized with the BD Cytofiix/Cytoperm Kit according to the manufacturers' instructions, and then stained 30 min with anti-IFN-γ (clone XMG1.2, eBiosciences), washed and analyzed by flow cytometry. Consensus B epitopes were deconvoluted to individual 15mers from peptide pools, where each peptide is present in two independent pools within the matrix and reactive peptides confirmed in the second round against the 15mer peptide. Finally, based on the sequence of the reactive 15mer peptide, truncated versions of the 15mer peptides were synthesized and tested.

T cell epitope mapping by ELISPOT assay. 14 days after immunization, splenocytes prepared (as detailed above) were re-stimulated *in vitro* with media alone, or with peptides, either in pools or individually (each at 1μM final concentration) derived from CN54 or ZM96 Gag on mouse anti-INF-γ antibody coated 96 well plates (U-Cytech) and incubated for 16 h at 37°C, 5% CO_2. IFN-γ production was revealed according to the manufacturer's instructions and IFN-γ spot forming cells (SFCs) enumerated using an immunospot image analyser (Bioreader 5000). In the first round, CN54 Gag epitopes were deconvoluted to individual 15mers from peptide pools, where each peptide is present in two independent pools within the matrix and reactive peptides confirmed in the second round against the 15mer peptide. Finally, based on the sequence of the reactive 15mer peptide, 9mer peptides were synthesized and tested. For ZM96 (due to the absence of a complete set of overlapping 15mer peptides), 49 individual 20mer peptides were tested. The reactive peptide sequences were confirmed against the corresponding 15mer peptide to the reactive sequence and then 9mer peptides synthesized and tested.

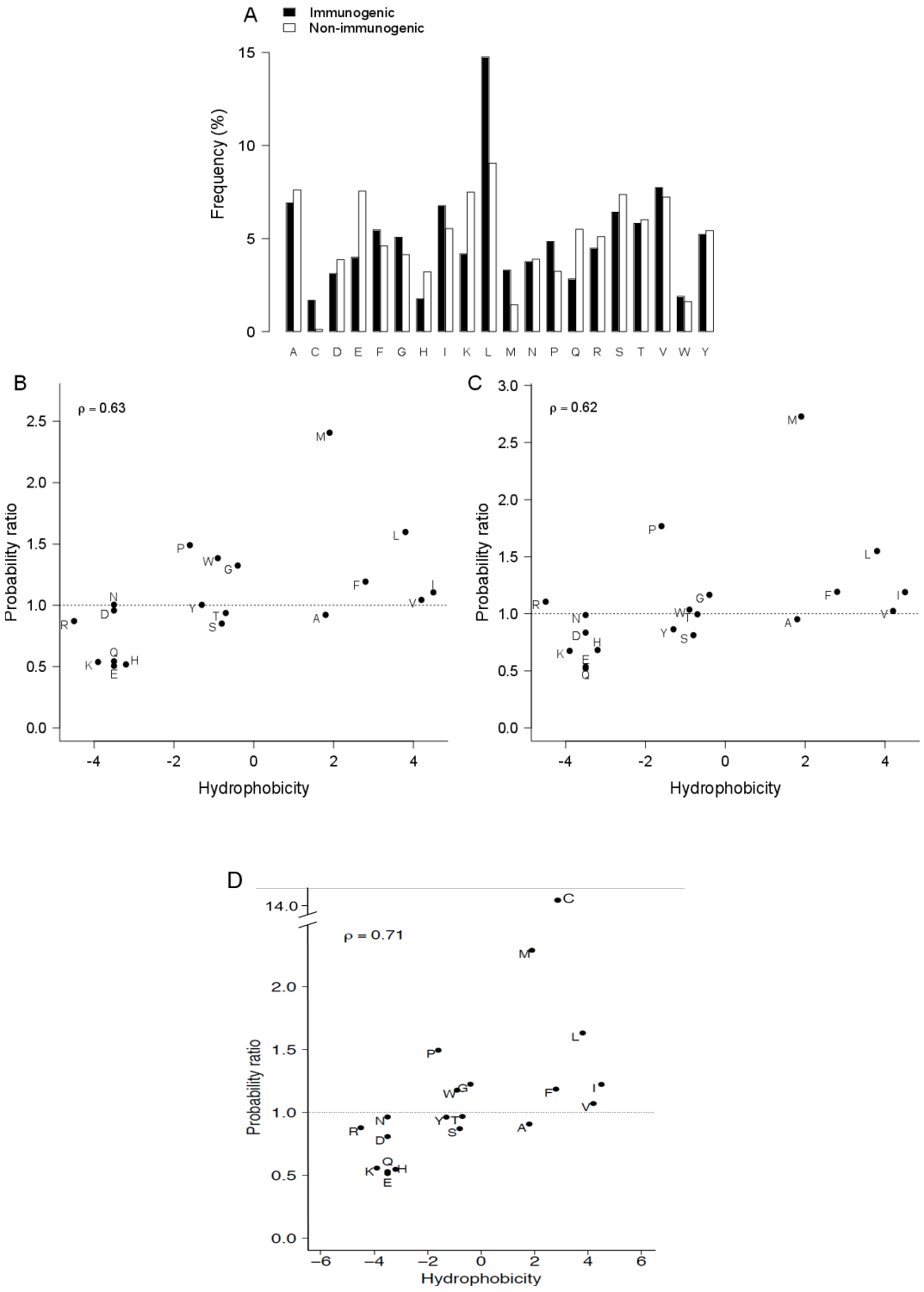

Fig. S1. Bias in amino acid usage between immunogenic and non-immunogenic MHC-I peptides. (*A*) Comparison of frequency distributions of amino acids between immunogenic and non-immunogenic datasets. (*B*) Probability ratio (P(x I immunogenic)/P(x I non-immunogenic)) of each amino acid as a function of its hydrophobicity, analyzed on just 9mer MHC-I peptides. (*C*) Probability ratio (P(x I immunogenic)/P(x I non-immunogenic)) of each amino acid as a function of its hydrophobicity, analyzed on 9mer HLA-I peptides excluding HLA-A2 restricted peptides. (*D*) Probability ratio of each amino acid as a function of its hydrophobicity; the plot shows cysteine (C) as an outlier in the immunogenic dataset.

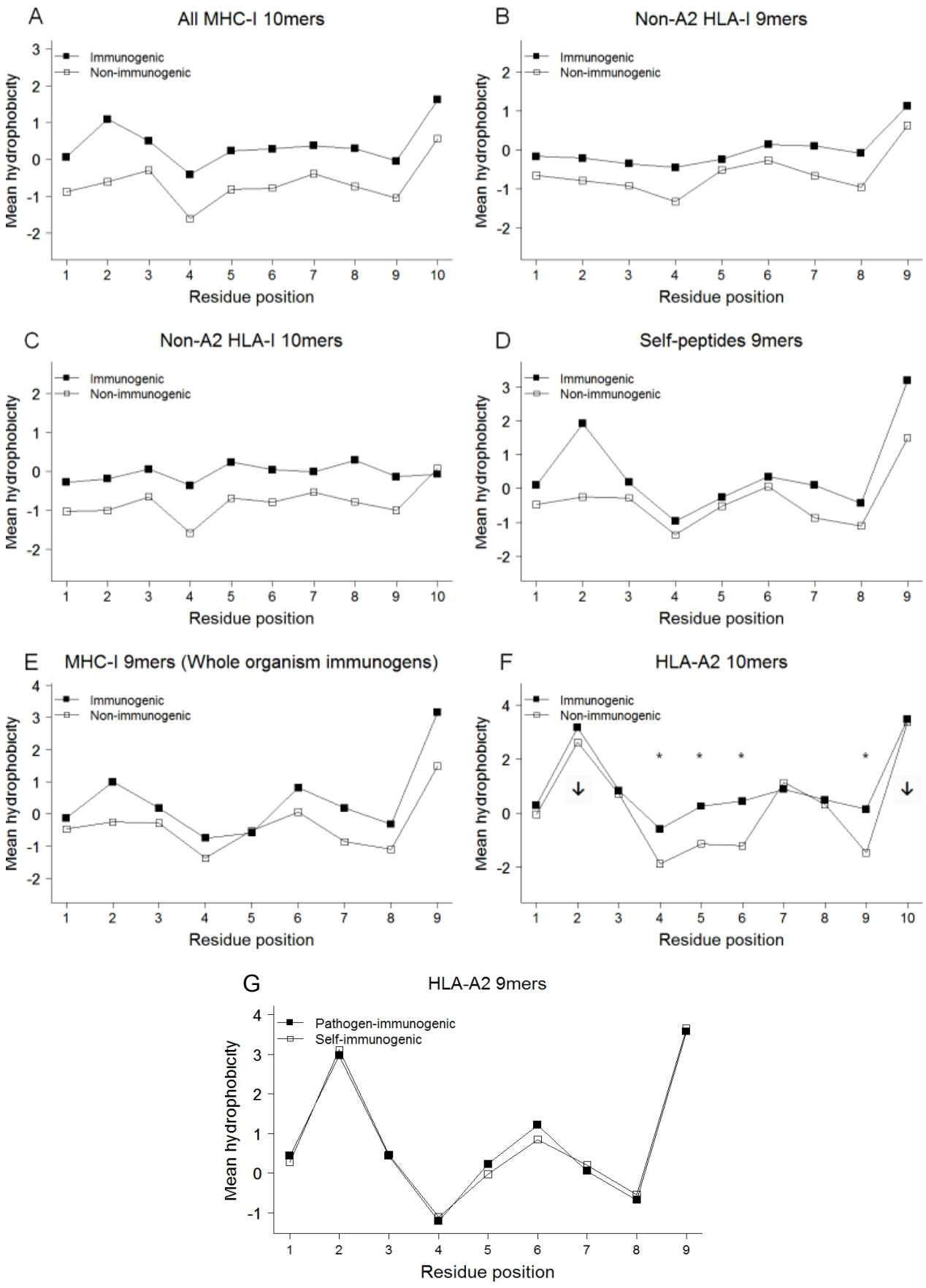
A
All MHC-I 10mers
Immunogenic
Non-immunogenic
Mean hydrophobicity
Residue position
B
Non-A2 HLA-I 9mers
C
Non-A2 HLA-I 10mers
D
Self-peptides 9mers
E
MHC-I 9mers (Whole organism immunogens)
F
HLA-A2 10mers
G
HLA-A2 9mers
Pathogen-immunogenic
Self-immunogenic

Fig. S2. Hydrophobicity comparison at each residue position between immunogenic and non-immunogenic MHC-I peptides, and immunogenic pathogen-derived and immunogenic self-epitopes. Each peptide sequence in the dataset was transformed into a numeric sequence based on amino acid hydrophobicity and the mean hydrophobicity at each position was computed. Unless indicated, analyzed peptides were not restricted to any MHC-motifs. (*A*) All immunogenic and non-immunogenic MHC-I 10mers; every single residue has $P < 2x10^{-7}$. (*B*) Human HLA-I immunogenic and non-immunogenic 9mers excluding HLA-A2 restricted peptides. (*C*) Human HLA-I immunogenic and non-immunogenic 10mers excluding HLA-A2 restricted peptides. (*D*) Immunogenic and non-immunogenic MHC-I 9mer self-peptides. (*E*) MHC-I 9mers peptides discovered using whole organism as immunogen as opposed to peptide-immunization experiments (non-immunogenic dataset – same as Fig. 2A) (*F*) Human HLA-A2 restricted immunogenic and non-immunogenic 10mers with arrows indicating anchor residues and stars for $P < 0.005$. (*G*) Human HLA-A2 restricted immunogenic pathogen-derived and immunogenic self 9mer epitopes. *P*-values for each figure were obtained using Wilcoxon rank-sum test and are shown in Table S3.

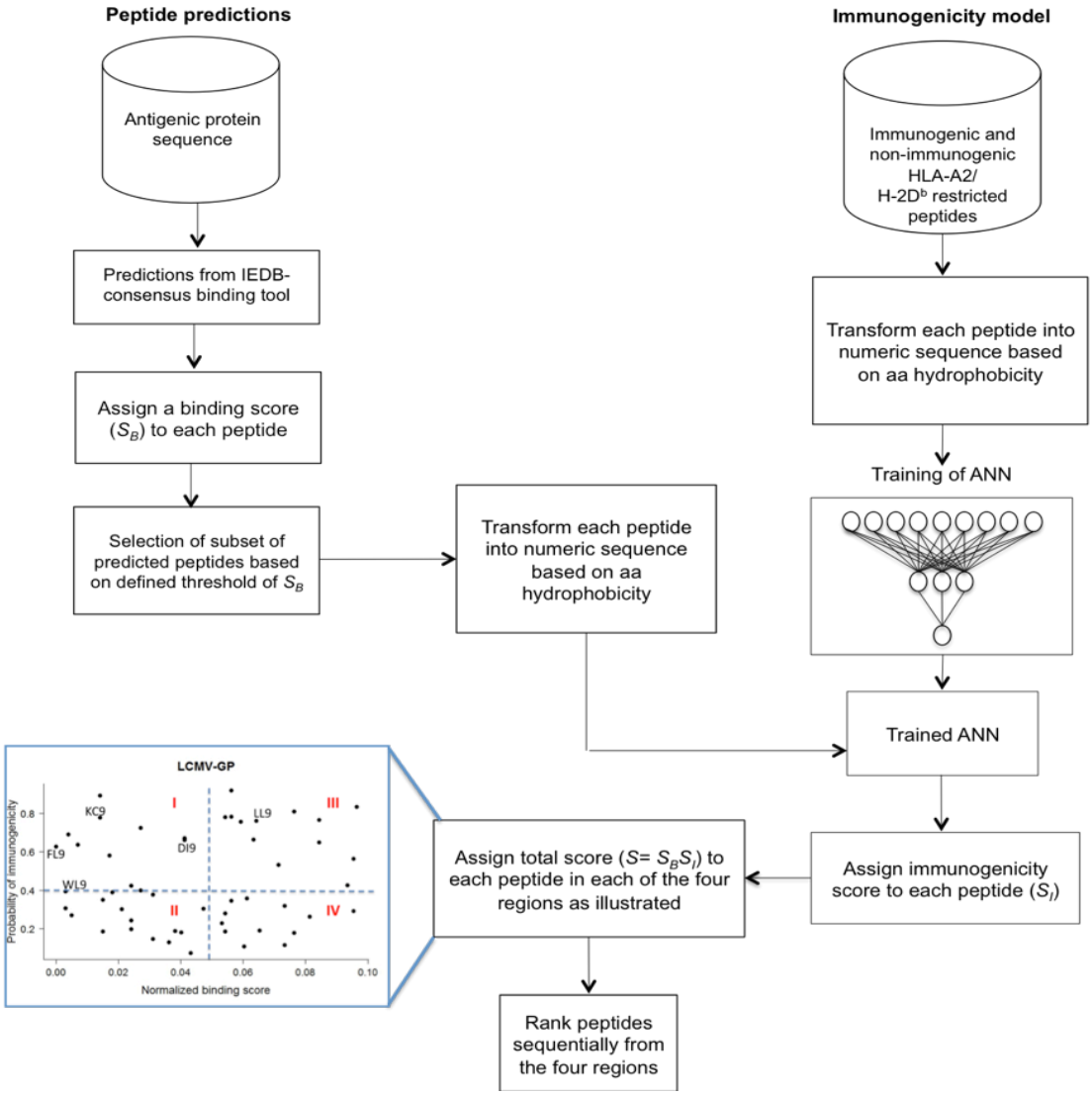

Fig. S3. Workflow for CTL epitope prediction using the ANN-Hydro model and the MHC-binding prediction tool IEDB-consensus. For training and application of the ANN-Hydro model for immunogenicity scores, each peptide sequence in the HLA-A2 and H-2D^b dataset was transformed into a corresponding numeric sequence based on the hydrophobicity value of amino acids. To obtain a list of candidates for MHC-bound peptides from a given antigen, IEDB-consensus binding algorithm was used and a normalized binding score (S_B) was assigned. The trained immunogenicity ANN model was applied on the same list of peptides independently to assign immunogenicity scores (S_I). After the subset of top binding peptides was selected, peptides from each region ranging from high-binding highly-immunogenic peptides to modest-binding low-immunogenic peptides (quadrants 1 through 4 in inset) were re-ranked based on total score $S = S_B \cdot S_I$. An example of epitope prediction is shown in the plot for experimentally defined H-2D^b restricted CTL epitopes from LCMV-GP. See Materials and Methods section Application of ANN-Hydro for full details.

	A	B	C	D	E	F	G	H	I	J	K	L
M	7872	7873	7874	7875	7876	7877	7878	7879	7880	7881	7882	7883
N	7884	7885	7886	7887	7888	7889	7890	7891	7892	7893	7894	7895
O	7896	7897	7898	7899	7900	7901	7902	7903	7904	7905	7906	7907
P	7908	7909	7910	7911	7912	7913	7914	7915	7916	7917	7918	7919
R	7920	7921	7922	7923	7924	7925	7926	7927	7928	7929	7930	7931
S	7932	7933	7934	7935	7936	7937	7938	7939	7940	7941	7942	7943
T	7944	7945	7946	7947	7948	7949	7950	7951	7952	7953	7954	7955
U	7956	7957	7958	7959	7960	7961	7962	7963	7964	7965	7966	7967
V	7968	7969	7970	7971	7972	7973	7974	7975	7976	7977	7978	7979
W	7980	7981	7982	7983	7984	7985	7986	7987	7988	7989	7990	7991
X	7992	7993	7994									

	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
P1	1	2	3	4	5	6	7	8	9	10	11
P2	12	13	14	15	16	17	18	19	20	21	22
P3	23	24	25	26	27	28	29	30	31	32	33
P4	34	35	36	37	38	39	40	41	42	43	44
P5	45	46	47	48	49	50	51	52	53	54	55
P6	56	57	58	59	60	61	62	63	64	65	66
P7	67	68	69	70	71	72	73	74	75	76	77
P8	78	79	80	81	82	83	84	85	86	87	88
P9	89	90	91	92	93	94	95	96	97	98	99
P10	100	101	102	103	104	105	106	107	108	109	110
P11	111	112	113	114	115	116	117	118	119	120	121

Fig. S4. Schematic of ConsB (top) and CN54 (bottom) 15mer peptide pools. Peptides were combined at 1μM/each peptide such that each peptide occurs in only two pools numbered 7872 –7994 for ConsB (top) or 7080.01-7080.121 for CN54 (indicated by 1–121, bottom). Yellow highlight indicates positive response to peptide pool. Green highlight indicates positive response to individual 15mer peptide, and red indicates negative response to individual 15mer peptide.

Source	Peptide	Amino Acid Sequence	Immunogenicity	Confirmed Epitope
Cons B	7889	QTGSEEL**RSLYNTVA**	+	H-2D^b
Cons B	7890	EEL**RSLYNTVAT**LYC	+	Gag76-84
CN54	7080.19	EEL**RSLFNTVAT**PYC	+	**RSLYNTVAT**
CN54	7080.20	**SLFNTVATPY**CVHTE	+	**RSLFNTVAT**
ZM96	7116.8	GTEEL**RSLYNTVAT**LYCVHE	+	**(RT9)**
Cons B	7912	EKAFSSPE**VIPMFSAL**	+	H-2K^b
Cons B	7913	SPE**VIPMFSAL**SEGA	+	Gag168-175
CN54	7080.41	F-SPE**VIPMFSAL**SEG	+	**VIPMFSAL**
CN54	7080.42	**VIPMFSAL**SEGATPQ	+	**VIPMFTAL**
ZM96	7116.17	EKAF-SPE**VIPMFTAL**SEGAT	+	**(VL8)**
Cons B	7920	GHQA**AMQMLKETI**NE	+	H-2D^b Gag197-205
Cons B	7921	**AMQMLKETI**NEEAAE	+	**AMQMLKETI**
CN54	7080.49	A**AMQILKDTI**NEEAA	+	**AMQILKDTI**
ZM96	7116.2	VGGHQA**AMQMLKDTI**NEEAA	+	**AMQMLKDTI (AI9)**
Cons B	7940	IVRM**YSPTSILDI**RQ	+	H-2D^b
Cons B	7941	**YSPTSILDI**RQGPKE	+	Gag277-285
CN54	7080.68	KIVRM**YSPTSILDI**K	+	**YSPTSILDI**
CN54	7080.69	M**YSPTSILDI**KQGPK	+	**YSPVSILDI**
ZM96	7116.28	NKIVRM**YSPVSILDI**KQGPK	+	**(YI9)**
Cons B	7949	ASQ**E**VKNWMTETLLV	–	H-2D^b
CN54	7080.77	QAT**QGVKNWMTDTL**L	+	Gag310-320
CN54	7080.78	**GVKNWMTDTL**LVQNA	+	**QGVKNWMTDTL**
ZM96	7116.32	Q**E**VKNWMTDTLLVQNA	–	**(QL11)**
Cons B	7963	EAM**SQVTNSATI**MMQ	+	H-2D^b
CN54	7080.91	AEAMSQ-TNSA-ILMQR	–	Gag368-376
ZM96	7116.37	RVLAEAMSQ-TNSVNILMQK	–	**SQVTNSATI (SI9)**

Fig. S5. Summary of identified epitopes. Responses to the RT9, VL8, AI9, and YI9 epitopes were observed for all three Gag protein variants, despite minor substitutions in the peptides. Overlapping sequences of individual peptides are shown. The QL11 epitope was only immunogenic for the CN54 Gag protein, but not ConsB or ZM96 Gag proteins, likely due to the A to E substitution at position 2. The SI9 epitope was only immunogenic in the ConsB Gag protein, as both CN54 and ZM96 had major deletions and substitutions in this sequence. MHC restriction was confirmed using MHC class I tetramer staining, and Gag amino acid positions are in reference to the HXB2 strain.

Amino acid	Hydrophobicity	Bulkiness	Polarity
Alanine (A)	1.8	11.5	8
Cysteine (C)	2.5	13.46	5.5
Aspartic acid (D)	-3.5	11.68	13
Glutamic acid (E)	-3.5	13.57	12.3
Phenylalanine (F)	2.8	19.8	5.2
Glycine (G)	-0.4	3.4	9
Histidine (H)	-3.2	13.69	10.4
Isoleucine (I)	4.5	21.4	5.2
Lysine (K)	-3.9	15.71	11.3
Leucine (L)	3.8	21.4	4.9
Methionine (M)	1.9	16.25	5.7
Asparagine (N)	-3.5	12.82	11.6
Proline (P)	-1.6	17.43	8
Glutamine (Q)	-3.5	14.45	10.5
Arginine(R)	-4.5	14.28	10.5
Serine (S)	-0.8	9.47	9.2
Threonine (T)	-0.7	15.77	8.6
Valine (V)	4.2	21.57	5.9
Tryptophan (W)	-0.9	21.67	5.4
Tyrosine (Y)	-1.3	18.03	6.2

Table S2. Amino acid property scales used for analyses. Hydrophobicity scale (Kyte-Doolittle) (5), Polarity (Grantham) (6), and Bulkiness (Zimmerman) (7).

Table S3.

Dataset	Sample size		Residue position in peptide									
	Immunogenic	Non Immunogenic	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
All MHC-I 9mers	2841	2660	6.6×10^{-15}	$< 2.2\times10^{-16}$	1.7×10^{-8}	$< 2.2\times10^{-16}$	4.6×10^{-5}	1.6×10^{-5}	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	-
All MHC-I 10mers	1007	1013	1.4×10^{-13}	$< 2.2\times10^{-16}$	1.7×10^{-9}	$< 2.2\times10^{-16}$	1.3×10^{-13}	1.7×10^{-15}	2×10^{-7}	3.3×10^{-12}	1.2×10^{-15}	$< 2.2\times10^{-16}$
NonA2 HLA-I 9mers	1318	1897	1.4×10^{-6}	7.8×10^{-6}	7.6×10^{-8}	$< 2.2\times10^{-16}$	0.1561	3×10^{-4}	2.6×10^{-13}	$< 2.2\times10^{-16}$	3.5×10^{-14}	-
NonA2 HLA-I 10mers	518	859	3.4×10^{-7}	1.4×10^{-6}	1.4×10^{-6}	4.6×10^{-13}	8.8×10^{-8}	1.4×10^{-6}	0.0023	1.5×10^{-8}	2.9×10^{-8}	0.2293
Self-peptides 9mers	565	2660	2.1×10^{-4}	$< 2.2\times10^{-16}$	1×10^{-4}	7.8×10^{-5}	0.292	0.1423	1.2×10^{-11}	2.2×10^{-7}	$< 2.2\times10^{-16}$	-
MHC-I 9mers (Immunogen=Organism)	323	2660	0.06781	6.5×10^{-9}	0.0071	1×10^{-5}	0.6156	1.3×10^{-4}	1.5×10^{-8}	3.2×10^{-6}	$< 2.2\times10^{-16}$	-
HLA A2 9mers	936	321	0.6384	0.8929	0.8487	6.3×10^{-12}	0.2917	3.1×10^{-7}	5×10^{-13}	$< 2.2\times10^{-16}$	0.08106	-
HLA A2 10mers	288	77	0.4029	0.2247	0.5054	0.00094	0.00087	1×10^{-5}	0.3368	0.9912	5.5×10^{-6}	0.09147
MHC H-2K^b 8mers	306	315	0.06678	0.9329	0.0055	0.2288	0.6742	7×10^{-5}	1.1×10^{-6}	0.1513	-	-
MHC H-2D^b 9mers	194	228	0.1825	0.0087	2.1×10^{-4}	0.3298	4.9×10^{-10}	0.7829	1.1×10^{-4}	0.0013	0.1273	-
	Immunogenic pathogenic	Immunogenic self										
HLA A2 9mers	1508	244	0.389	0.385	0.877	0.97	0.13	0.04	0.776	0.35	0.63	

Anchor residues
TCR contact residues

Hydrophobicity comparison between immunogenic and non-immunogenic MHC class I peptides. *P*-values were calculated using Wilcoxon sum-ranked test using different datasets as described in main text.

Antigen	Epitope	Epitope length
CMV-pp65	NLVPMVATV	9
	MLNIPSINV	9
	VLGPISGHV	9
	RLLQTGIHV	9
	LMNGQQIFL	9
	ILARNLVPM	9
	SLILVSQYT	9
	SIYVYALPL	9
	VIGDQYVKV	9
	YLESFCEDV	9
	AMAGASTSA	9
	KYQEFFWDA	9
	GLSISGNLL	9
	RQYDPVAAL	9
	VAALFFFDI	9
	ALFFFDIDL	9
	KISHIMLDVA	10
	SDNEIHNPAV	10
	FTWPPWQAGI	10
	LLCPKSIPGL	10
Dengue-Polyprotein	VLMLVAHYA	9
	ILLMRTTWA	9
	MLLALIAVL	9
	TLYAVATTI	9
	QEGAMHTAL	9
	LPAIVREAI	9
	SRNSTHEMY	9
	AIVREAIKR	9
	YLPAIVREA	9
	TLLCLIPTV	9
	VLNPYMPSV	9
	LMMMLPATL	9
	VTYECPLLV	9
	MMMLPATLA	9
	IILEFFLMV	9
	KTDFGFYQV	9
	VQADMGCVV	9
	GLLFMILTV	9
	QLWAALLSL	9
	LLMRTTWAL	9
	CLMMMLPATL	10
	ELMRRGDLPV	10
	MLLILCVTQV	10
	FLMVLLIPEP	10
	TLMLLALIAV	10
	LMLLALIAVL	10
	IILEFFLMVL	10
	TLTAAVLLLV	10
	VLLLVTHYAI	10
	ITLLCLIPTV	10
	KVLNPYMPSV	10
	HQLWATLLSL	10
	YTPEGIIPTL	10
	SIILEFFLMV	10
	LSMGLITIAV	10
	NQLIYVILTI	10
	LMMMLPATLA	10
	TLMAMDLGEL	10
	FTMGVLCLAI	10

Table S4. HLA-A2 restricted CTL epitopes for dengue virus 1 polyprotein and cytomegalovirus pp65 used in the rate analysis of predicted epitopes as shown in Fig. 3. All epitopes were retrieved from IEDB (1).

Source	Epitope	Antigen	p-ANN-Hydro	Reference
Neo-epitopes				
Rotavirus	SLISGMWLL	Rota-VP2_4	0.89	Newell et al.
	TLLANVTAV	Rota-VP6_4	0.87	
	FLDSEPHLL	Rota-NSP1_2	0.84	
	LLNYILKSV	Rota-VP7_1	0.37	
Influenza-A (FluA)	QIAILVTTV	NA	0.90	Assarsson et al.
	GLIYNRMGA	M1	0.89	
	GILGFVFTL	Flu_1	0.80	
	FVEALARSI	PB1	0.58	
	VMNILLQYL	GAD	0.57	
	FVANFSMEL	PB1	0.54	
	TTYQRTRAL	NP	0.47	
	GLADQLIHL	HIV_7	0.47	
Dengue-Virus 2 (DENV-2)	GLLTVCYVL	NS2B	0.88	Weiskopf et al.
	RLITVNPIV	E	0.88	
	IMAVGMVSI	NS2B	0.85	
	IILEFFLIV	NS4A	0.79	
	ALSELPETL	NS4A	0.61	
	YLPAIVREA	NS3	0.50	
	KLAEAIFKL	NS5	0.44	
	AAAWYLWEV	NS3	0.27	
Positive Control epitopes				
Human herpesvirus 5 (CMV)	ALFFFDIDL	CMV_5	0.84	Newell et al.
	LMNGQQIFL	CMV_18	0.81	
	RIFAELEGV	CMV_22	0.81	
	QMWQARLTV	CMV_21	0.78	
	NLVPMVATV	CMV_1	0.75	
	VLEETSVML	CMV-IE1	0.72	
	FLMEHTMPV	CMV_8	0.61	
	IIYTRNHEV	CMV_13	0.54	
	SLLSEFCRV	CMV_23	0.52	
	ILSPLTKGI	CMV_15	0.46	
	VLAELVKQI	CMV_2	0.24	
Human herpesvirus 4 (EBV)	GLCTLVAML	EBV_2	0.87	Newell et al.
	YVLDHLIVV	EBV_1	0.84	
	YLQQNWWTL	EBV_5	0.79	
	CLGGLLTMV	EBV_4	0.79	
	YLLEMLWRL	EBV_3	0.28	
Influenza-A (FluA)	FLDIWTYNA	Flu_4	0.85	
	NMLSTVLGV	Flu_14	0.78	
	LLIDGTASL	Flu_12	0.68	

	FMYSDFHFI	Flu_5	0.63	
	MMMGMFNML	Flu_13	0.48	
	GMFNMLSTV	Flu_7	0.29	
	RLIDFLKDV	Flu_15	0.23	
Hepatitis B virus (HBV)	WLSLLVPFV	HBV_2	0.89	Newell et al.
	FLLSLGIHL	HBV_5	0.74	
	FLLTRILTI	HBV_1	0.70	
Human Immunodeficiency virus (HIV-1)	NVWATHACV	HIV_1	0.94	
	TLNAWVKVV	HIV_2	0.86	
	KLTPLCVTL	HIV_4	0.84	
	SLYNTVATL	HIV_5	0.61	
	ALVEMGHHA	HIV_8	0.44	
	ILKEPVHGV	HIV_9	0.38	
	LTFGWCFKL	HIV_6	0.36	
Mycobacterium *tuberculosis*	GLPVEYLQV	TB_1	0.85	
	KLIANNTRV	TB	0.79	
Plasmodium *falciparum*	YLNKIQNSL	CSP	0.79	
LCMV	YLVSIFLHL	LCMV	0.77	
Herpes simplex virus (HSV-1)	SLPITVYYA	HSV1/2	0.90	
RSV	KMLKEMGEV	RSV	0.29	
Self Antigens	ALWMRLLPL	pp-Insulin	0.83	
	YMCSFLFNL	EZH2	0.51	
	YMDGTMSQV	Tyrosinase	0.49	

Table S5. Probabilities assigned by the ANN-Hydro 'A2-model' for HLA-A2 restricted 9mer CTL epitopes. Three recent epitope discovery studies (8–10) that were based on a proteome-wide screen of various viral antigens and self-epitopes were chosen for assessment of the predictive capacity of the 'A2-model'. Neo-epitopes were obtained from rotavirus (10) dengue virus (9) and influenza A (8) and other positive control epitopes from several antigens (pathogenic and self) were obtained from Newell et al. (10). Any epitope that was present in the training set for ANN-Hydro was removed. A cutoff probability (p-ANN-Hydro) of 0.4 was set for a positive "hit".

Antigen	Epitope	MHC	Length	IEDB bind	Syfpe-ithi	NetMHC .bind	IEDB. prot	ANN .prot	ANN Hydro + IEDB.bind	p-ANN-Hydro	Reference
LCMV-GP	FALISFLLL	H-2D^b	9	1	10	1	3	1	1	0.62	(11)
	WLVTNGSYL		9	3	1	2	5	4	11	0.4	(12)
	LIDYNKAAL		9	45	12	39	68	77	32	0.8	(13)
	KAVYNFATC		9/11	8	9	10	39	53	5	0.77	(11, 14)
	DEVINIVII		9	24	4	74	133	115	10	0.66	(13)
LCMV-NP	FQPQNGQFI		9	1	1	1	2	1	1	0.69	(14)
	SEVSNVQRI		9	7	2	7	37	50	12	0.14	(13)
Ad.v.T. antigen	VNIRNCCYI		9	1	1	1	1	1	13	0.31	(15)
	CSDGNCHLL		9	21	4	9	11	44	20	0.8	(15)
Flu-NP	ASNENMETM		9	1	1	1	1	2	1	0.87	(16)
	RLIQNSLTI		9	3	3	2	3	3	2	0.67	(16)
	GERQNATEI		9	18	2	36	100	103	8	0.42	(16)
	YRRVNGKWM		9	19	4	35	80	65	9	0.44	(16)
FluA-Neuraminidase	FCGVNSDTV		9	3	2	4	11	3	13	0.35	(17)
	ITYKNSTWV		9	4	8	3	4	2	1	0.52	(17)
	YRYGNGVWI		9	5	7	11	29	7	4	0.45	(17, 18)
Consensus Gag	SQVTNSATI		9	1	1	1	1	1	7	0.2	This study
	AMQMLKETI		9	4	9	6	4	2	9	0.19	
	YSPTSILDI		9	6	19	4	11	3	11	0.39	
	RSLYNTVAT		9	3	32	5	45	29	1	0.82	
ZM96 Gag	AMQMLKDTI		9	1	4	4	3	2	13	0.17	This study
	YSPVSILDI		9	5	12	3	9	1	3	0.43	
	RSLYNTVAT		9	4	28	5	46	27	2	0.77	
97CN54 Gag	AMQILKDTI		9	1	4	4	3	2	13	0.16	This study
	YSPTSILDI		9	5	19	2	9	3	15	0.36	
	RSLFNTVAT		9	2	35	3	40	23	2	0.76	
Melan-A	ALMDKSLHV	A2	9	1	3	1	1	1	1	0.65	(19)
	GILTVILGV		9	2	1	2	2	2	2	0.86	(19)
	ILTVILGVL		9	7	5	5	4	5	8	0.9	(19)
	AAGIGILTV		9/10	9	4	7	19	17	9	0.86	(19)
Wt-1	SLGEQQYSV		9	1	3	3	3	3	1	0.79	(20)
	RMFPNAPYL		9	2	10	2	2	2	2	0.53	(20)
	ALLPAVPSL		9	3	1	1	1	1	3	0.44	(20)
	DLNALLPAV		9	6	2	8	16	27	4	0.9	(20)
	VLDFAPPGA		9	7	31	7	15	13	8	0.87	(20)
	KLGAAEASA		9	9	16	9	21	26	5	0.94	(20)
	NLGATLKGV		9	12	4	10	12	14	12	0.76	(20)
	CMTWNQMNL		9	13	27	11	7	9	13	0.69	(20)
	RVPGVAPTL		9	25	17	21	11	11	19	0.81	(20)
gp100	RLMKQDFSV		9	1	21	1	2	2	1	0.83	(21)

	MLGTHTMEV	A2	9	2	15	2	3	1	2	0.65	(21)
	KTWGQYWQV		9	5	62	3	4	3	5	0.5	(21)
	YLEPGPVTA		9	16	20	19	34	25	11	0.93	(21)
TRAG-3	GLIQLVEGV		9	1	1	1	1	1	1	0.43	(22, 23)
	HACWPAFTV		9	9	10	6	9	20	9	0.82	(22, 23)
	SILLRDAGL		9	6	2	8	5	7	5	0.92	(22, 23)
	ILLRDAGLV		9	3	3	2	4	4	3	0.7	(22, 23)
	ALSKFPRQL		9	4	5	4	3	2	4	0.34	(22, 23)
p53	RMPEAAPPV		9	1	1	8	2	2	1	0.51	(24)
	LLGRNSFEV		9	2	2	4	1	4	2	0.46	(24)
	VVPCEPPEV		9	13	14	21	14	11	9	0.77	(25)
	YQGSYGFRL		9	5	5	65	3	3	10	0.41	(24)
	KTCPVQLWV		9	14	15	19	34	25	11	0.77	(24)

Table S6. Ranking comparison of all the predicted epitopes (Prevalidation and *in vivo* validation) used in this study as shown in Fig 5A. The predictions used are as follows: ANN-Hydro - ANN-hydrophobicity prediction model combined with normalized binding scores from prediction algorithms, IEDB-Bind -IEDB consensus binding tool, NetMHC-Bind - NetMHCpan binding tool, SYFPEITHI - SYFPEITHI epitope prediction tool, IEDB-Prot-IEDB recommended processing prediction, ANN-Prot - IEDB processing predictions using ANN. p-ANN-Hydro – Probability of immunogenicity assigned by the corresponding (H-2D^b or A2) ANN-Hydro immunogenicity model. All references were obtained from IEDB (1) and are indicated.

ConsB Gag predictions				
Rank	**Epitope**	**Binding score *(S_B)***	**p-ANN-Hydro**	**Total score *(S)***
1	**RSLYNTVAT**	**0.006**	**0.82**	**0.001**
2	ATPQDLNTM	0.033	0.77	0.008
3	QVSQNYPIV	0.049	0.83	0.008
4	RFAVNPGLL	0.025	0.66	0.008
5	RMYSPTSIL	0.039	0.72	0.011
6	KARVLAEAM	0.021	0.40	0.013
7	**SQVTNSATI**	**0.000**	**0.20**	**0.000**
8	SQVSQNYPI	0.004	0.12	0.004
9	**AMQMLKETI**	**0.008**	**0.19**	**0.006**
10	GWMTNNPPI	0.008	0.14	0.007
11	**YSPTSILDI**	**0.020**	**0.39**	**0.012**
12	RSLFGNDPS	0.023	0.27	0.017
13	ASVLSGGEL	0.022	0.07	0.020
14	KALGPAATL	0.036	0.36	0.023
15	AAMQMLKET	0.039	0.33	0.026
16	VQNANPDCK	0.053	0.84	0.009
17	SALSEGATP	0.054	0.84	0.009
18	LLVQNANPD	0.051	0.76	0.012
19	ASLRSLFGN	0.096	0.79	0.020
20	SLYNTVATL	0.061	0.64	0.022

ZM96 Gag predictions				
Rank	**Epitope**	**Binding score *(S_B)***	**p-ANN-Hydro**	**Total score *(S)***
1	MSQTNSVNI	0.000	0.64	0.000
2	**RSLYNTVAT**	**0.001**	**0.78**	**0.000**
3	**YSPVSILDI**	**0.004**	**0.43**	**0.002**
4	YMIKHLVWA	0.016	0.77	0.004
5	KVSQNYPIV	0.032	0.83	0.005
6	ATPQDLNTM	0.028	0.75	0.007
7	RMYSPVSIL	0.040	0.80	0.008
8	VQNANPDCK	0.048	0.83	0.008
9	LLVQNANPD	0.046	0.81	0.009
10	RFALNPGLL	0.027	0.67	0.009
11	KARVLAEAM	0.016	0.42	0.009
12	NFLQNRPEP	0.042	0.61	0.017
13	**AMQMLKDTI**	**0.000**	**0.18**	**0.000**
14	KSLFGSDPL	0.000	0.08	0.000
15	KALGPGATL	0.026	0.36	0.017
16	KIVRMYSPV	0.026	0.18	0.021
17	IMKQLQPAL	0.039	0.34	0.026
18	VKNWMTDTL	0.033	0.18	0.027
19	AWMTSNPPI	0.033	0.10	0.030
20	WMTSNPPIP	0.092	0.92	0.007

CN54 Gag predictions				
Rank	**Epitope**	**Binding score *(S_B)***	**p-ANN-Hydro**	**Total score (*S*)**
1	MSQTNSAIL	0.002	0.85	0.0003
2	**RSLFNTVAT**	**0.002**	**0.76**	**0.0004**
3	YMLKHLVWA	0.018	0.72	0.005
4	KVSQNYPIV	0.032	0.83	0.005
5	ATPQDLNTM	0.028	0.78	0.006
6	SALSEGATP	0.049	0.84	0.008
7	RFALNPGLL	0.027	0.70	0.008
8	VQNANPDCK	0.048	0.82	0.008
9	LLVQNANPD	0.046	0.78	0.010
10	RMYSPTSIL	0.034	0.70	0.010
11	NFLQNRPEP	0.042	0.65	0.015
12	SALQTGTEE	0.042	0.57	0.018
13	**AMQILKDTI**	**0.000**	**0.16**	**0.000**
14	KAKVLAEAM	0.012	0.35	0.008
15	**YSPTSILDI**	**0.015**	**0.36**	**0.010**
16	RALGPGASI	0.020	0.24	0.015
17	KSLFGNDPS	0.025	0.28	0.018
18	IMKQLQSAL	0.037	0.33	0.025
19	VKNWMTDTL	0.033	0.20	0.027
20	WMTSNPPVP	0.077	0.92	0.006

Table S7. Ranked list of top 20 predicted peptides for each of the Gag variant using the ANN-Hydro combined with normalized binding scores from predictions. (S_B) - Binding score, p-ANN-Hydro - probability of immunogenicity obtained by applying ANN-Hydro model to each peptide, (*S*) - Total score. This list was ranked based on total score *S* ranging from lowest score to the highest score within each section (I through IV) classified based on p-ANN-Hydro and S_B (see section Application of ANN-Hydro).

Supporting References

1. Vita R, et al. (2010) The immune epitope database 2.0. *Nucleic Acids Res* 38(Database issue):D854–D862. Available at: www.iedb.org [Accessed October 20, 2013].

2. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (ISBN 3-900051-07-0). Available at: http://www.r-project.org.

3. Rammensee H, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S (2000) SYFPEITHI : database for MHC ligands and peptide motifs. *Immunogenetics* (1999):213–219.

4. Bachy V, et al. (2013) Langerin negative dendritic cells promote potent CD8+ T-cell priming by skin delivery of live adenovirus vaccine microneedle arrays. *Proc Natl Acad Sci U S A* 110(8):3041–3046.

5. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157(1):105–132.

6. Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* 185(4154):862–864.

7. Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 21(2):170–201.

8. Assarsson E, et al. (2008) Immunomic analysis of the repertoire of T-cell specificities for influenza A virus in humans. *J Virol* 82(24):12241–51.

9. Weiskopf D, et al. (2011) Insights into HLA-restricted T cell responses in a novel mouse model of dengue virus infection point toward new implications for vaccine design. *J Immunol* 187(8):4268–79.

10. Newell EW, et al. (2013) Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat Biotechnol* 31(7):623–629.

11. Kotturi MF, et al. (2007) The CD8+ T-cell response to lymphocytic choriomeningitis virus involves the L antigen: uncovering new tricks for an old virus. *J Virol* 81(10):4928–4940.

12. Hudrisier D, Riond J, Mazarguil H, Gairin JE (2001) Pleiotropic effects of post-translational modifications on the fate of viral glycopeptides as cytotoxic T cell epitopes. *J Biol Chem* 276(41):38255–38260.

13. Oldstone MB, et al. (1999) Use of a high-affinity peptide that aborts MHC-restricted cytotoxic T lymphocyte activity against multiple viruses in vitro and virus-induced immunopathologic disease in vivo. *Virology* 256(2):246–257.

14. Van der Most RG, et al. (1998) Identification of Db- and Kb-restricted subdominant cytotoxic T-cell responses in lymphocytic choriomeningitis virus-infected mice. *Virology* 240(1):158–67.

15. Toes REM, et al. (1995) An Adenovirus Type 5 Early Region I B-Encoded CTL Epitope-Mediating Tumor Eradication by CTL Clones Is Down-Modulated by an Activated ras Oncogene. *J Immunol*:3396–3405.

16. Oukka M, et al. (1996) Protection Against Lethal Viral Infection by Vaccination with Nonimmunodominant Peptides. *J Immunol*:3039–3045.

17. Zanker D, Waithman J, Yewdell JW, Chen W (2013) Mixed proteasomes function to increase viral peptide diversity and broaden antiviral CD8+ T cell responses. *J Immunol* 191(1):52–59.

18. Zhong W, Reche P a, Lai C-C, Reinhold B, Reinherz EL (2003) Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. *J Biol Chem* 278(46):45135–45144.

19. Elsas A Van, Burg SH Van Der, Borghi M, Mourer JS (1996) Peptide-pulsed dendritic cells induce tumoricidal cytotoxic T lymphocytes from healthy donors against stably HLA-A*0201-binding peptides from the Melan-A / MART-1 self antigen. *Eur J Immunol* 26:1683–1689.

20. Doubrovina E, et al. (2012) Mapping of novel peptides of WT-1 and presenting HLA alleles that induce epitope-specific HLA-restricted T cells with cytotoxic activity against WT-1(+) leukemias. *Blood* 120(8):1633–1646.

21. Tsai V, et al. (1997) Identification of Subdominant CTL Epitopes of the GP100 Melanoma-associated Tumor Antigen by Primary In Vitro Immunization with Peptide-pulsed Dendritic Cells. *J Immunol* 22:1796–1802.

22. Meier A, et al. (2005) Spontaneous T-cell responses against peptides derived from the Taxol resistance-associated gene-3 (TRAG-3) protein in cancer patients. *Cancer Immunol Immunother* 54(3):219–228.

23. Zhu B, et al. (2003) Identification of HLA-A * 0201-restricted Cytotoxic T Lymphocyte Epitope from TRAG-3 Antigen. *Clin cancer Res* 9(May):1850–1857.

24. Svane IM, et al. (2004) Vaccination with p53-peptide-pulsed dendritic cells, of patients with advanced breast cancer: report from a phase I study. *Cancer Immunol Immunother* 53(7):633–641.

25. Ito D, et al. (2007) Immunological characterization of missense mutations occurring within cytotoxic T cell-defined p53 epitopes in HLA-A*0201+ squamous cell carcinomas of the head and neck. *Int J Cancer* 120(12):2618–2624.