Computational MHC-I epitope predictor identifies 95% of experimentally mapped HIV-1 clade A and D epitopes in a Ugandan cohort.

CURRENT STATUS: ACCEPTED

BMC Infectious Diseases  ▪  BMC Series

Daniel Bugembe Lule  dan.lule@mrcuganda.org
MRC / UVRI & LSHTM Uganda Research Unit
Corresponding Author
ORCiD: 0000-0002-5616-515X

Andrew Obuku Ekii
MRC/UVRI Uganda Research Unit On Aids

Christine Watera
Uganda Virus Research Institute

Nicaise Ndembí
Institute of Human Virology, Abuja, Nigeria

Jennifer Serwanga
MRC/UVRI & LSHTM Uganda Research Unit

Pontiano Kaleebu
MRC Lifecourse Epidemiology Unit

Pietro Pala
MRC/UVRI Uganda Research Unit On Aids

DOI:
10.21203/rs.2.14495/v2

SUBJECT AREAS
Infectious Diseases

KEYWORDS
HIV-1, epitope mapping, T-cell, artificial neural network, in-silico, NetMHCpan4.0, MHCflurry1.2.0 and NetCTL1.2.
Abstract

Background: Identifying immunogens that induce HIV-1-specific immune responses is a lengthy process that can benefit from computational methods, which predict T-cell epitopes for various HLA types. Methods: We tested the performance of the NetMHCpan4.0 computational neural network in re-identifying 93 T-cell epitopes that had been previously independently mapped using the whole proteome IFN-g ELISPOT assays in 6 HLA class I typed Ugandan individuals infected with HIV-1 subtypes A1 and D. To provide a benchmark we compared the predictions for NetMHCpan4.0 to MHCflurry1.2.0 and NetCTL1.2. Results: NetMHCpan4.0 performed best correctly predicting 88 of the 93 experimentally mapped epitopes for a set length of 9-mer and matched HLA class I alleles. Receiver Operator Characteristic (ROC) analysis gave an area under the curve (AUC) of 0.928. Setting NetMHCpan4.0 to predict 11-14mer length did not improve the prediction (37-79 of 93 peptides) with an inverse correlation between the number of predictions and length set. Late time point peptides were significantly stronger binders than early peptides (Wilcoxon signed rank test: p =0.0000005). MHCflurry1.2.0 similarly predicted all but 2 of the peptides that NetMHCpan4.0 predicted and NetCTL1.2 predicted only 14 of the 93 experimental peptides. Conclusion: NetMHCpan4.0 class I epitope predictions covered 95% of the epitope responses identified in six HIV-1 infected individuals, and would have reduced the number of experimental confirmatory tests by >80%. Algorithmic epitope prediction in conjunction with HLA allele frequency information can cost-effectively assist immunogen design through minimizing the experimental effort. Keywords: HIV-1, epitope mapping, T-cell, artificial neural network, in-silico, NetMHCpan4.0, MHCflurry1.2.0 and NetCTL1.2.

Background
Computational algorithms are increasingly utilised in biological modelling and offer the potential to reduce the time and expense of immunological assays. Computational algorithms were initially demonstrated as useful tools for predicting potential epitopes that might elicit quality T-cell responses [1, 2]. Computational algorithms that predict potential HLA binding T-cell epitopes can facilitate the design of vaccines capable of inducing T-cell immunity against HIV-1. The high variability of HIV-1 and the extensive genetic polymorphism of HLA molecules can be managed in silico, allowing immunogen optimisation to increase breadth and magnitude of T cell responses in respect of HLA allele frequencies and circulating virus strains in different populations. Bioinformatics approaches were previously applied as proof of concept for an HIV-1 peptide-based vaccine for the env and gag genes [3] in cynomolgus macaques for a broad spectrum of HIV-1 clades. Computational optimisation of immunogens facilitates the development of the multivalent and mosaic vaccines [4] necessary to control recombinant HIV-1 strains, an increasingly common occurrence in the epidemic in Uganda [5]. Computational approaches aim to identify optimal epitopes relevant to vaccine development and are not isolated to HIV-1 only, but a wide range of pathogens, including Ebola virus [6], therefore various statistical validation approaches have been applied for evaluation of these methods [7-10].

For HIV-1 vaccine design purposes an important consideration for the suitability of a computational algorithm is the breadth of discrete number of T-cell epitopes it generates that could reach particular levels of coverage [11] of circulating viruses. The higher the number of epitope variants the more the reduction in their requirements to attain optimum coverage levels for any epidemic. Previous data that has shown that breadth of T-cell response is associated to viral set point in chronic HIV-1 infection [12-17]. In order to translate the computational epitope prediction into vaccine design, the number of
discrete epitopes computationally generated from particular HIV-1 proteins is an important metric for further investigation [11].

A reliable pan-HLA-specific algorithm NetMHCpan4.0 [18-20] that has been improved by advances in HLA binding data, covers 172 MHC class I molecules from human (HLA-A, B, C, E), mouse (H-2), cattle (BoLA), primates (Patr, Mamu, Gogo) and swine (SLA) [20, 21], and can also predict binding to alleles devoid of experimental data basing on similarity to known binders and non-binders [22, 23]. This is an artificial neural network (ANN) algorithm for predictions of 8-14aa and capable of predicting epitopes for other HLA alleles using data for similar alleles by positional similarity of residues in their binding motifs. NetMHCpan4.0 is considered to be the tool of choice for such predictions considering the benchmarking done against other related tools [24]. Nevertheless to have a conclusive outcome of the computational performance we compared NetMHCpan4.0 to both an older and recent tool, NetCTL1.2 [25-27] and MHCflurry1.2.0 [28] respectively.

The binding of CTL epitopes to MHC class I molecules is linear, anchoring at residues 2 and 9; hence the interface between ligand and CTL can be determined computationally [29]. Validation of such computational applications can be done by comparing their predictions with suitable experimental data. Despite the paucity of data validating the performance of computational methods relative to wet laboratory experiments, a few have documented them to achieve an area under the curve AUC of over 90% [18, 19, 30, 31] by isolated experimental data. We have not come across a wet experiment that evaluated computational predictors to achieve a robust AUC using a single set of wet laboratory experimental data. The previously reported 90% AUC is largely based on positional specific scoring algorithms (PSSM) for the collective isolated experiments alongside probability models used to establish affinity or binding scores. One study that explored the reliability of in-silico approaches in epitope prediction and its application for vaccine
design reported a meagre 22%, 44%, and relatively higher 78% match for three computational tools namely YFPEITHI, CTLPRED and IEDB respectively [32]. Using experimental epitope mapping data generated from 757 peptides tested on cells of 6 early HIV-1 infected individuals at paired time points, we show that NetMHCpan4.0 can be useful for markedly reducing pooled peptide experiments as demonstrated by the 95% experimental and computational concordance.

Methods

Experimental Binder Data

The data used was from an independent study that did not include this analysis in its objectives. Experimental data of peptides previously mapped for HIV-1 epitope recognition of 6 individuals for a separate study (Table 4) at 2 time points each was used for comparison with the positive computationally predicted binders. These were from a Ugandan early HIV-1 serodiscordant couple cohort approved by the Uganda Virus Research Institute (UVRI) Research and Ethics review board and the Uganda National Council of Science and Technology (UNCST). All participants provided informed consent. Six (6) participants whose experimental epitope recognition profile we evaluated were early HIV-1 infections (Table 4), enrolled under the following criteria: (i) detection of HIV-1 P24 antigen with a simultaneous negative HIV-1 antibody ELISA (2 participants) or documented HIV-1 sero-negative test in the previous 12 months (4 participants); (ii) HAART naïve (all). Early infection was determined following the Fiebig Staging criteria[33] as described elsewhere by Obuku A.E. et.al.[34].

The experimentally tested peptides totalled 757 (figure 5), were 17aa long, overlapping by 11aa and spanning the HIV-1 proteome consensus for subtypes A and D. Cultured ELISpot assays using 200000 cells/well as previously documented by Obuku AE. et.al. [34] and ex-vivo IFN-g ELISpot assay using 100000 cells/well were used for testing peptide pools and
epitope mapping respectively. Experimental positive pools were 3 times the background wells and at least 600 spot forming units per million cells. “Deconvolute This” software [35] was used to identify possible responding individual peptides from the pools or where it was not possible all the peptides in a pool were tested as single peptides (figure 6).

**HLA typing**

High resolution reference strand conformation analysis HLA class I tissue typing for the early infected subjects was done using methods described elsewhere [36].

**HIV-1 subtyping**

HIV-1 subtyping determination was performed on the *gag* gene [37, 38] using Sanger method generated sequences. The sequences were input into the REGA HIV-1 automated subtyping tool to determine the HIV-1 clade [39, 40].

**Computation Epitope Prediction.**

The HIV-1 A and D subtype consensus sequences were used as inputs for the computational epitope prediction. These peptide sequences were all for the year 2004 downloaded from the Los Alamos database (hiv.lanl.gov/content/sequence/NEWALIGN/align.html). The web version of NetMHCpan4.0 [19] (http://www.cbs.dtu.dk/services/NetMHCpan/) was configured to predict 9mer through 14mer epitopes for 22 HLA class I alleles (Table 4) that were expressed by the 6 HIV-1 infected donors. Linux version MHCflurry1.2.0 [28] was used to predict 9mer epitopes and an earlier tool NetCTL1.2 was also used to predict 9mer epitopes for the 22 HLA class I alleles expressed by the 6 study individuals. Perl version 5.26.2 was used to extract the binders from all the NetMHCpan4.0 predictions and also to compare the computational binders to the 93 mapped experimental 17aa peptides for 9mer through 14mer hits using a sliding window. An experimental peptide was considered a hit if any of the computational 9mer through 14mer sequence was contained in the 17 amino acid
experimental peptide sequence as well as any of the HLA-A, B or C expressed by the individual matched the NetMHCpan4.0 HLA class I type(s). If multiple computational epitope predictions were contained in a single 17mer experimental peptide they counted as a single hit. These were determined by a BLAST search of the computational binders against the derivative experimental peptides to determine computational predictions from the same test peptide. The accession numbers of the sequences used to determine the HIV-1 subtypes for 5 of the 6 study subjects are; KT825896, KT825897, KT825898, KT825899, KT825900, KT825901, KT825902, KT825903, KT825904, KT825905, KT825906, KT825907, KT825908, KT825909, KT825910, KT825911 and KT82512.

**Data Analysis**

Statistics computations and plots were generated using SPSS version 24.0.0.0. The NetMHCpan4.0 computational performance was evaluated using a confusion matrix to classify true positives, true negatives, false positives and false negatives that were used for the Receiver Operator Characteristic (ROC) plot. The hit rate (sensitivity) and false hit rate (specificity) of binder predictions as determined by the NetMHCpan4.0 threshold of peptides within the top 2% (with a score of 2 or less) were calculated and the strength of the model was determined by calculating the area under the curve, AUC of the ROC plot [41-43]. Pearson’s correlation coefficient was used to evaluate the relationship between the number of epitopes with various HIV-1 genes. To evaluate if there were any differences in the early versus late time point peptides for the binding ranking of the experimentally mapped peptides as predicted by the computational score the Wilcoxon signed rank test was used. To evaluate if HIV-1 subtypes A and D affected the number of computational predictions generated, Fisher Exact Test was used. To determine whether multiple computationally predicted epitope sequences were derived from the same experimental peptide sequence, a local blast database was set up using Geneious version.
9.0.5. Both clades A and D experimental consensus sequences were used separately each as a reference sequence for the blast. The computational peptide sequences were then aligned against the consensuses to evaluate those derived from a single 17 amino acid experimental peptide sequence. Where an experimental peptide was predicted by multiple or overlapping computational peptides, the average NetMHCpan4.0 score was assigned as the computational score for this peptide. This score was also used during the generation of the ROC curve and the confusion matrix. To compare the association between ELISPOT spot forming units and NetMHCpan4.0 scores or MHCflurry1.2.0 affinities and also the association between the values for the 2 computational tools, Pearson’s correlation coefficient was used.

Results

**Number of Experimental Assays Compared to Computationally Guided Prediction**

**Assay Projections**

To experimentally determine epitopes for 757 peptides spanning the whole HIV-1 proteome for clades A and D as well as both time points of the 6 individuals required a total of 4230 test assay wells. For each test subject these included 9 antigen proliferation wells, 384 culture ELISPOT wells and an average of 164 epitope mapping ELISPOT wells (Range; 148 – 186 test wells). Using the 22 HLA alleles represented study subjects we were able to computationally predict 95% of the experimentally mapped epitopes. This approach could have reduced the test assays by eliminating all the T-cell antigen proliferation and culture ELISPOT steps totalling to 3258 assay wells (77%) and leaving only 972 (23%) epitope mapping assay required. Applying a pooling strategy to the computational predictions similar to that used in the experimental pooling where each pool contained approximately 20 peptides with a coverage of 3 per peptide, the 923 potential peptides (95% of experimental peptides for epitope mapping ELISPOT derived
from the 972 (23%) eligible epitope mapping peptides) would make at most 46 pools.
Consequently the computational prediction approach could have reduced the experimental
assays by at least 80%.

**Magnitude of Epitope Predictions are Variable across HLA Alleles, HIV-1 Proteins and Clades**

The input HIV-1 subtypes A1 and D consensus whole proteome sequences evaluated for potential 9, 10, 11, 12, 13 and 14-mer binders to the 22 HLA alleles represented in the six patients, varied in the distribution of predicted binders across HIV-1 genes and HLA alleles. All the peptide hits predicted for 10 through 14-mer were also all predicted in the 9-mer set except for two 14-mer peptides. An expected positive correlation for protein size with number of predictions as well as with HIV-1 protein length was observed as illustrated by Spearman’s rank order correlation; $r_s=0.88$ (Figures 1, A and B).

NetMHCpan4.0 predicted 95% (88/93) of the experimentally mapped peptides as binders and missed 5% (5 out of 93) for the 12-time points of the 6 participants (Table 2).

MHCflurry predicted 91% (85/93) of the experimental peptides and had a lot of similarity to NetMHCpan4.0 for the predicted HLA. NetCTL was the least performing tool with only 15% (14/93) predicted experimental peptides (Table 2).

Comparison of the various epitope prediction length set showed that the 9mer setting was ideal for NetMHCpan4.0. The number of predictions were 88, 79, 55, 39, 39 and 37 hits out of 93 for 9, 10, 11, 12, 13 and 14-mer epitopes respectively. Increasing the prediction length from 9mer through 14mer resulted in a smaller number of predicted binders as illustrated in Figure 2. Since we held the assumption that our wet experimental data was the gold standard we evaluated the sensitivity and specificity of the NetMHCpan4.0. The computational predictor had more predicted binders than those determined by the experimental mapping as presented in the confusion matrix in Table 1. The experimental
positive’s count also shown in table 2 under column “Hits” shows the test peptide count (1 through 88) that contained the computational 9-mer sequence. Multiple computational epitopes may be contained in a single experimental peptide, as shown in the column “NetMHCpan4.0 9-mer Epitope Prediction” in Table 2. Overall HIV-1 Clade A 9-mer predictions were fewer in number than clade D (figure 1, C) though the difference did not approach statistical significance.

Comparison of Experimentally Mapped Epitopes with in-silico Prediction

The experimental peptide mapping data was derived from a baseline time point corresponding to HIV-1 Fiebig stages IV, V and VI (table 4) and a later time point. Ninety-three (n=93) epitopes were experimentally mapped of which 12 were recognized at both baseline and later time points, 34 only at baseline and 54 only at the later time point. Comparison of the ranked computational score for Netmhcpan4.0 binders of early (n=34) versus later peptides showed that the later time point predictions were stronger binders reaching statistical significance (Wilcoxon signed rank p-value=0.0000005) (figure 3). NetMHCpan4.0 ranked binders as those predicted to be in the top 2% and assigned a score of 0.2 or below. Any binder within the top 0.5% and assigned a score of 0.05 or below was ranked as a strong binder. Considering only the 9-mer computational predictions, peptides that were derived from the same 17-mer experimental peptide were determined by a BLAST mapping to their derivative sequences. The 17-mer peptides were then classified into a confusion matrix (table 1) as either true positives, false positive, true negative and false negative. From the classification the true positive rate (sensitivity) was plotted against the false positive rate (1-specificity) for only 9-mer predictions using an ROC curve and the AUC attained reached 0.928 (Figure 4). Only 9-mer length epitopes were considered in the ROC analysis as increasing the length to 10-mer through 14mer NetMHCpan4.0 predictions neither raised the number of predicted binders nor improved
the hit rate as all their predictions contained the sequence already predicted in the 9-mer set except 1 14-mer peptide (hit 72 in table 2).

Comparison of the ELISPOT magnitude of response (spot forming units) did not show any association to either NetMHCpan4.0 scores or MHCflurry1.2.0 affinity values. Similarly a comparison of the latter 2 computational predictors did not show any association between their assigned “affinity” values.

We observed a negative correlation for the number of computationally predicted epitopes with the length of input for netMHCpan4.0 (Figure 2). NetMHCpan4.0 registered the highest concordance to the wet experiments followed by MHCflurry1.2.0.

Discussion

In this analysis we showed that the computational method NetMHCpan4.0 predicted 95% of previously experimentally mapped HIV-1 epitopes in 6 HIV-1 infected individuals expressing a total of 22 different HLA class I alleles. In our IFN-g ELISPOT assays we evaluated 757 17mer peptides overlapping by 11 amino acids and covering the whole subtype A1 and D consensus proteomes. Out of the 5 experimentally determined epitopes missed by the algorithm, 4 were actually computationally predicted as binders but were not included for lack of concordance with the participant’s HLA alleles. About one third (37) of 125 total positive predictions were not experimentally supported in our tests. These do not necessarily represent false positives, as ELISPOT detection depends on the frequency of specific T cells in the participant’s repertoire, and we observed changes in dominant T cell specificities within a given participant between early and later time points after HIV-1 infection. A formal ROC evaluation of the score generated by NetMHCpan4.0 as a classifier for peptides recognised/not recognised by PBMC in IFN-g ELISPOT assays, produced an AUC of 0.928. Thus experimental confirmatory tests cannot be dropped altogether, however the NetMHCpan4.0 algorithm could provide a considerable saving of
time and resources in verifying just the predicted epitopes.
As participants had been enrolled in the acute/early phase of HIV-1 infection and we had observed intra-participant changes in epitope recognition between early and late time points after infection, we compared the binding scores of confirmed epitopes at these time points and found a statistically significant change towards recognition of higher binding peptides as the infection entered the chronic phase. This might represent better support of the T-cell response directed at more stable HLA/peptide complexes as the infection progresses into chronicity.

The NetMHCpan4.0 algorithm, which is based on binding affinity and integrates data on eluted naturally processed ligands, reflected optimal HLA class I binding for 9-mers, producing a decreasing number of predictions when the peptide size was increased from 9 to 11 amino acids, with minor differences for further increases between sizes of 11, 12, 13 and 14 amino acids. With a single exception, predicted binders between 11-14 amino acids included at least one 9mer predicted to bind on its own, suggesting that a destabilizing effect of the extra amino acids beyond the canonical HLA class I binding pockets at positions 2 and 9 could account for fewer predictions.

Important limitations are the lack of predictions of HLA class II restricted epitopes, which might have contributed to a fraction of IFN-g ELISPOT responses. Approximately 5% of the computational predictions may be false positives that only increase the size of planned wet experiments and approximately 1% of true positives may also be missed.

Conclusion
In this analysis, using NetMHCpan4.0, MHCflurry and NetCTL to predict previously experimentally mapped epitopes, we demonstrate that the computational methods reliably predict an acceptable portion of binder epitopes. We recommend the use of such computational methods to reduce the size of experiments required cost associated.
List Of Abbreviations

AUC - Area under the curve

CTL – Cytotoxic T lymphocytes

HIV-1 – Human immunodeficiency virus type 1

HLA – Human leucocyte antigen

ROC – Receiver operator characteristic

Declarations

**Ethics approval and consent to participate.**

The Rubicon study, from which we derived the experimental data, was reviewed and approved by the Uganda Virus Research Institute, Research and Ethics Committee (UVRI-REC) and the Uganda National Council of Science and Technology (UNCT). All participants provided informed signed consent accepting to freely participate in this study.

**Consent for publication.**

Not applicable.

**Availability of data and materials.**

Most of the relevant data to support the manuscript has been included in the write-up. If any addition data is required will be availed once requested.

**Competing interests.**

The authors declare that they have no competing interests.

**Funding.**

This research is jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement, the Wellcome Trust (grant WT078927MA), and EDCTP (project code: TA_05_40200_40203).
Author's contributions.

DBL performed the peptide mapping experiments. CW recruited participants. NN provided HIV-1 subtyping. DBL analysed the data together with PP. DBL wrote the manuscript with contributions from PP, AEO and JS. All authors reviewed the manuscript and or provided useful contributions as well as approved the final manuscript.

Acknowledgement.

We thank the study participants who provided specimen for the wet laboratory experiments, the AIDS Information Centre Clinic Kampala, Uganda that steered the Rubicon discordant couple cohort study and the late Anthony Kebba who initiated the Rubicon Cohort.

We thank Ruhena Sargeant for HLA typing, the late Harr F. Njai for HIV-1 ELISA and Western blot assays and Deogratius Ssemwanga for help with GenBank submissions.

References

1. Tambunan US, Sipahutar FR, Parikesit AA, Kerami D: Vaccine Design for H5N1 Based on B- and T-cell Epitope Predictions. *Bioinformatics and biology insights* 2016, 10:27-35.

2. Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V: Methods for prediction of peptide binding to MHC molecules: a comparative study. *Molecular medicine* 2002, 8(3):137-148.

3. Azizi A, Anderson DE, Torres JV, Ogrel A, Ghorbani M, Soare C, Sandstrom P, Fournier J, Diaz-Mitoma F: Induction of broad cross-subtype-specific HIV-1 immune responses by a novel multivalent HIV-1 peptide vaccine in cynomolgus macaques. *Journal of immunology* 2008, 180(4):2174-2186.

4. Baden LR, Walsh SR, Seaman MS, Cohen YZ, Johnson JA, Licona JH, Filter RD, Kleinjan JA, Gothing JA, Jennings J et al: First-in-Human Randomized Controlled Trial of
Mosaic HIV-1 Immunogens Delivered via a Modified Vaccinia Ankara Vector. J Infect Dis 2018.

5. Yebra G, Ragonnet-Cronin M, Ssemwanga D, Parry CM, Logue CH, Cane PA, Kaleebu P, Brown AJ: Analysis of the history and spread of HIV-1 in Uganda using phylodynamics. J Gen Virol 2015, 96(Pt 7):1890-1898.

6. Lim WC, Khan AM: Mapping HLA-A2, -A3 and -B7 supertype-restricted T-cell epitopes in the ebolavirus proteome. BMC genomics 2018, 19(Suppl 1):42.

7. Goodswen SJ, Kennedy PJ, Ellis JT: Enhancing in silico protein-based vaccine discovery for eukaryotic pathogens using predicted peptide-MHC binding and peptide conservation scores. PloS one 2014, 9(12):e115745.

8. Goodswen SJ, Kennedy PJ, Ellis JT: Discovering a vaccine against neosporosis using computers: is it feasible? Trends in parasitology 2014, 30(8):401-411.

9. Goodswen SJ, Kennedy PJ, Ellis JT: Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. Bioinformatics 2014, 30(16):2381-2383.

10. Pedersen LE, Rasmussen M, Harndahl M, Nielsen M, Buus S, Jungersen G: A combined prediction strategy increases identification of peptides bound with high affinity and stability to porcine MHC class I molecules SLA-1*04:01, SLA-2*04:01, and SLA-3*04:01. Immunogenetics 2016, 68(2):157-165.

11. Currier JR, Robb ML, Michael NL, Marovich MA: Defining epitope coverage requirements for T cell-based HIV vaccines: theoretical considerations and practical applications. Journal of translational medicine 2011, 9:212.

12. Frahm N, Kiepiela P, Adams S, Linde CH, Hewitt HS, Sango K, Feeney ME, Addo MM, Lichterfeld M, Lahaie MP et al: Control of human immunodeficiency virus replication by cytotoxic T lymphocytes targeting subdominant epitopes.
13. Frahm N, Korber BT, Adams CM, Szinger JJ, Draenert R, Addo MM, Feeney ME, Yusim K, Sango K, Brown NV et al: **Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities.** *Journal of virology* 2004, **78**(5):2187-2200.

14. Geldmacher C, Currier JR, Herrmann E, Haule A, Kuta E, McCutchan F, Njovu L, Geis S, Hoffmann O, Maboko L et al: **CD8 T-cell recognition of multiple epitopes within specific Gag regions is associated with maintenance of a low steady-state viremia in human immunodeficiency virus type 1-seropositive patients.** *Journal of virology* 2007, **81**(5):2440-2448.

15. Goonetilleke N, Moore S, Dally L, Winstone N, Cebere I, Mahmoud A, Pinheiro S, Gillespie G, Brown D, Loach V et al: **Induction of multifunctional human immunodeficiency virus type 1 (HIV-1)-specific T cells capable of proliferation in healthy subjects by using a prime-boost regimen of DNA- and modified vaccinia virus Ankara-vectored vaccines expressing HIV-1 Gag coupled to CD8+ T-cell epitopes.** *Journal of virology* 2006, **80**(10):4717-4728.

16. Liu Y, McNevin JP, Holte S, McElrath MJ, Mullins JL: **Dynamics of viral evolution and CTL responses in HIV-1 infection.** *PloS one* 2011, **6**(1):e15639.

17. Rolland M, Heckerman D, Deng W, Rousseau CM, Coovadia H, Bishop K, Goulder PJ, Walker BD, Brander C, Mullins JL: **Broad and Gag-biased HIV-1 epitope repertoires are associated with lower viral loads.** *PloS one* 2008, **3**(1):e1424.

18. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M: **NetMHCpan, a method for MHC class I binding prediction beyond humans.** *Immunogenetics* 2009, **61**(1):1-13.

19. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M: **NetMHCpan-4.0:**
Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology* 2017, 199(9):3360-3368.

20. Zhang H, Lundegaard C, Nielsen M: **Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods.** *Bioinformatics* 2009, 25(1):83-89.

21. Peters B, Bui HH, Frankild S, Nielsen M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W *et al.: A community resource benchmarking predictions of peptide binding to MHC-I molecules.** *PLoS computational biology* 2006, 2(6):e65.

22. Jacob L, Vert JP: **Efficient peptide-MHC-I binding prediction for alleles with few known binders.** *Bioinformatics* 2008, 24(3):358-366.

23. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V: **MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides.** *Nucleic acids research* 2005, 33(Web Server issue):W172-179.

24. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, Sette A, Peters B, Nielsen M: **Automated benchmarking of peptide-MHC class I binding predictions.** *Bioinformatics* 2015, 31(13):2174-2181.

25. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M: **An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions.** *European journal of immunology* 2005, 35(8):2295-2303.

26. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M: **Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction.** *BMC bioinformatics* 2007, 8:424.

27. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S,
Lund O: Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein science: a publication of the Protein Society* 2003, 12(5):1007-1017.

28. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J: MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell systems* 2018, 7(1):129-132 e124.

29. DeLisi C, Berzofsky JA: T-cell antigenic sites tend to be amphipathic structures. *Proceedings of the National Academy of Sciences of the United States of America* 1985, 82(20):7048-7052.

30. He Y, Rappuoli R, De Groot AS, Chen RT: Emerging vaccine informatics. *Journal of biomedicine & biotechnology* 2010, 2010:218590.

31. Reche PA, Reinherz EL: Prediction of peptide-MHC binding using profiles. *Methods in molecular biology* 2007, 409:185-200.

32. Roider J, Meissner T, Kraut F, Vollbrecht T, Stirner R, Bogner JR, Draenert R: Comparison of experimental fine-mapping to in silico prediction results of HIV-1 epitopes reveals ongoing need for mapping experiments. *Immunology* 2014, 143(2):193-201.

33. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, Heldebrant C, Smith R, Conrad A, Kleinman SH et al: Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS* 2003, 17(13):1871-1879.

34. Obuku AE, Bugembe DL, Musinguzi K, Watera C, Serwanga J, Ndewbi N, Levin J, Kaleebu P, Pala P: Macrophage Inflammatory Protein-1 Beta and Interferon Gamma Responses in Ugandans with HIV-1 Acute/Early Infections. *AIDS research and human retroviruses* 2016, 32(3):237-246.
35. Roederer M, Koup RA: **Optimized determination of T cell epitope responses.** *Journal of immunological methods* 2003, **274**(1-2):221-228.

36. Arguello JR, Little AM, Bohan E, Goldman JM, Marsh SG, Madrigal JA: **High resolution HLA class I typing by reference strand mediated conformation analysis (RSCA).** *Tissue antigens* 1998, **52**(1):57-66.

37. Serwanga J, Mugaba S, Pimego E, Nanteza B, Lyagoba F, Nakubulwa S, Heath L, Nsubuga RN, Ndemi N, Gotch F et al: **Profile of T cell recognition of HIV type 1 consensus group M Gag and Nef peptides in a clade A1- and D-infected Ugandan population.** *AIDS research and human retroviruses* 2012, **28**(4):384-392.

38. Serwanga J, Nakiboneka R, Mugaba S, Magambo B, Ndemi N, Gotch F, Kaleebu P: **Frequencies of Gag-restricted T-cell escape "footprints" differ across HIV-1 clades A1 and D chronically infected Ugandans irrespective of host HLA B alleles.** *Vaccine* 2015, **33**(14):1664-1672.

39. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, Galvao-Castro B, Vandamme AM, de Oliveira T: **A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences.** *Nucleic acids research* 2009, **37**(Web Server issue):W634-642.

40. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AM, van de Vijver DA et al: **An automated genotyping system for analysis of HIV-1 and other microbial sequences.** *Bioinformatics* 2005, **21**(19):3797-3800.

41. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143**(1):29-36.

42. Park SH, Goo JM, Jo CH: **Receiver operating characteristic (ROC) curve: practical review for radiologists.** *Korean journal of radiology* 2004, **5**(1):11-18.
Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER: Small-sample precision of ROC-related estimates. Bioinformatics 2010, 26(6):822-830.

Tables

Table 1. Experimental and computational 9mer peptide confusion matrix.

| Identifier | Participant's HLA Types | Hit No. | Screening Peptide | Screening Peptide HIV-1 Clade | True Positive (88) (Hits in table 2) | False Positive (37) | False Negative (5) | True Negative (627) |
|------------|--------------------------|---------|-------------------|------------------|----------------------------------------|----------------------|-------------------|-----------------------|
| E91        | A*02:01                  | 1       | FITKGLG15YGRKRRQR | D                | GLGISYGRK                             |                      |                   |                       |
|            | A*03:01                  |         |                   |                  |                                        |                      |                   |                       |
|            | B*53:01                  |         |                   |                  |                                        |                      |                   |                       |
|            | B*58:02                  |         |                   |                  |                                        |                      |                   |                       |
|            | Cw*04:01                 |         |                   |                  |                                        |                      |                   |                       |
|            | Cw*06:02                 |         |                   |                  |                                        |                      |                   |                       |
| E92        | A*02:01                  | 3       | RKQNPEIV15YQYMDDLYV | D                | YQYMDDLYV                             |                      |                   |                       |
|            | A*30:02                  |         |                   |                  |                                        |                      |                   |                       |
|            | B*44:03                  |         |                   |                  |                                        |                      |                   |                       |
|            | B*14:02                  |         |                   |                  |                                        |                      |                   |                       |
|            | Cw*04:01                 |         |                   |                  |                                        |                      |                   |                       |
|            | Cw*08:02                 |         |                   |                  |                                        |                      |                   |                       |
|            | ELNKRTQDFWEVQLGLPH       | A1      |                   |                  |                                        |                      |                   |                       |
|            | ELNKRTQDFWEVQLGLPH       | A1      |                   |                  |                                        |                      |                   |                       |
|            | ELNKRTQDFWEVQLGLPH       | A1      |                   |                  |                                        |                      |                   |                       |
|            | ELNKRTQDFWEVQLGLPH       | A1      |                   |                  |                                        |                      |                   |                       |
|            | NDIQKLVGKLNWASQ1YP       | D        |                   |                  |                                        |                      |                   |                       |
|            | NDIQKLVGKLNWASQ1YP       | D        |                   |                  |                                        |                      |                   |                       |
|            | NDIQKLVGKLNWASQ1YP       | D        |                   |                  |                                        |                      |                   |                       |
|            | PAIQTGSEELRSLNTVA        | D        |                   |                  |                                        |                      |                   |                       |
|            | PAIQTGSEELRSLNTVA        | D        |                   |                  |                                        |                      |                   |                       |
|            | PYNTPIFAIIKDKSTKW        | A1      |                   |                  |                                        |                      |                   |                       |
|            | GANNSHNETFRP6GGG1MR      | D        |                   |                  |                                        |                      |                   |                       |
|            | GMDGPVKQWPLTEEK1K        | A1      |                   |                  |                                        |                      |                   |                       |
|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 10 | PLTSKLFGNDPLSQ | D | KSLFGNDPL | A' | Cw |
|   | PLTSKLFGNDPLSQ | D | KSLFGNDPL |   |   |
|   | FKGPRIKICNCFCGEGHI | D |   |   |   |
|   | HERI DEVDKEAKL | D | EVKDTKAEAL | B' | B' |
|   | HERI DEVDKEAKL | D | IEVKDTKEA |   |   |
| E94 | A*34:02 | KIEQI ONSKQQK | A1 | EIONNSKOK | A' |
|   | A*74:01 | NHPSCVWLEAQEEEVEVGF | A1 | LEAQQEEEVE | B' |
|   | B*44:03 | HQDPIPKQPSSQPRGD | D | HQDPIPKQP | Cw |
|   | B*58:02 |   |   |   |   |
|   | Cw*04:01 |   |   |   |   |
|   | Cw*06:02 |   |   |   |   |
|   |   |   |   |   |   |
| E95 | A*23:01 | VAVHVASGIEA | A1 | VAVHVASG | Cv |
|   | A*74:01 | KRWIILGLNIVRMYSPV | A1 | WIILGLNKI | A* |
|   | B*44:03 | KRWIILGLNIVRMYSPV | A1 | ILGLNIVK | A* |
|   | B*15:10 | NMMLNIVGHHQAMQMLK | A1 | HQAAMQMLK | B* |
|   | Cw*04:01 | NMMLNIVGHHQAMQMLK | A1 | HQAAMQMLK | A' |
|   | Cw*16:01 | NMMLNIVGHHQAMQMLK | A1 | HQAAMQMLK | Cw |
|   |   |   |   |   |   |
|   |   |   |   |   |   |
| E913 | A*02:01 | LWQRPVLTIKIGGQLKEA | D | LWQRPVLTIT | A' |
|   | A*34:02 | LWQRPVLTIKIGGQLKEA | D | QRPVLTIKI | Cw |
|   | B*45:01 | LWQRPVLTIKIGGQLKEA | D | WQRPLVTIK | B' |
|   | B*47:01 | TVPVKLKPMDGPVKQW | A1 | LKPMDGPK | A* |
| Protein | Amino Acid Sequence | Antibody Type |
|---------|--------------------|---------------|
| E914    | DKWASLWNWFSITQWLWY | A*            |
|         | DKWASLWNWFSITQWLWY | Cw            |
|         | DKWASLWNWFSITQWLWY | A*            |
|         | PVDPEVEKATEGENNSL  | A*            |
| L91     | EQMHTDIISLWDQSLK   | A*            |
|         | EQMHTDIISLWDQSLK   | Cw            |
|         | EQMHTDIISLWDQSLK   | A*            |
|         | LETSEGCKIQIIGQLQPAI| A*            |
|         | LETTEGCQQIMEQQLQPAI| A*            |
|         | LETTEGCQQIMEQQLQPAI| Cw            |
|         | LETTEGCQQIMEQQLQPAI| A*            |
|         | ERILSTCLGRSAEPVPL  | A1            |
|         | ERILSTCLGRSAEPVPL  | A*            |
|         | ERILSTCLGRSAEPVPL  | B*            |
|         | LVGPTPVIIGRNMLTQI  | A*            |
|         | CKQIIGQLQPAIQTGSEEL| A*            |
|         | CKQIIGQLQPAIQTGSEEL| A*            |
|         | PAIQTGSEELRSLYNTVA | A*            |
|         | PAIQTGSEELRSLYNTVA | Cw            |
| L92     | NDIOQLVGKLNWASQIYP | A*            |
|         | NDIOQLVGKLNWASQIYP | A*            |
|         | NDIOQLVGKLNWASQIYP | Cw            |
|         | NDIOQLVGKLNWASQIYP | A*            |
|         | LVKTYWGLHTGEREWHL  | A*            |
|         | LVKTYWGLHTGEREWHL  | A*            |
|         | SLVNVRQGYSPLSFQTL  | D             |
|         | SLVNVRQGYSPLSFQTL  | B*            |
|         | SLVNVRQGYSPLSFQTL  | Cw            |
|         | SLVNVRQGYSPLSFQTL  | A*            |
|         | TLPCRIOJKINMQUOGY  | A*            |
|         | MRVRGQNYQHQLRWR    | B*            |
|         | GEMKNC5F8ITTEIRDCK | B*            |
|         | NVTENFMWKNMVEQMH   | Cw            |
|         | NVTENFMWKNMVEQMH   | B*            |
| L94     | LIHLHYDFCDSAIRKA   | A1            |
|         | LIHLHYDFCDSAIRKA   | Cw            |
|         | LIHLHYDFCDSAIRKA   | Cw            |
|         | LIHLHYDFCDSAIRKA   | A*            |
| B*5802 | LIHLHYFDCFSDSAIRKA | A1 | FSDSAIRKA | Cw |
|-------|-------------------|----|-----------|----|
|       | HLARNCRPRKKGGCWK  | A1 | ARNCRPRK  | A  |
|       | HLARNCRPRKKGGCWK  | A1 | HLARNCR   | A  |
|       | HLARNCRPRKKGGCWK  | A1 | HLARNCR   | A' |
| L95   | SKQKTQAAADTGNSSKV  | A1 | AADTGNSSK | A  |
|       | HQDPIPKQPSQPGRGD   | D  | HQDPIPKQP  | Cw |
| A*23:01 | KRWIILGLNKIVRMYSPV | A1 | WIILGLNKI | A* |
| A*74:01 | NMMLNIVGGHQAAMQMLK | A1 | GHQAAMQML | B' |
| B*44:03 | NMMLNIVGGHQAAMQMLK | A1 | HQAAMQMLK | A' |
|       | NMMLNIVGGHQAAMQMLK | A1 | HQAAMQMLK | Cw |
|      | EVNIVTDSQYALGIIQA  | A1 | EVNIVTDSQ | B' |
| A*34:01 | AAETEMHNWVATHACV   | A1 | TEMHNWVAT | B' |
| A*34:02 | AAEWDRLHPVHAGPI    | A1 | AAEWDRLP  | B' |
| A*74:01 | AAEWDRLHPVHAGPI    | A1 | LHPVHAGPI | B' |
| Cw*04:01 | AAEWDRLHPVHAGPI   | A1 | RLHPVHAGP | A' |
| Cw*16:01 | LRALGPGATLEEMMTA  | A1 | RALGPGATL | B' |
|      | FFKTLRAEQATQEVKNWM | A1 | AEQATQEVK | B' |
|      | GTTSTPOEQIQGWMTGNPPI | A1 | QEQIGWMTP | B' |
|      | WKVIEEKAFSPEVIPMF  | A1 | EKEFSPEV | B' |
|      | WKVIEEKAFSPEVIPMF  | A1 | WKVIEEKA  | A' |
|      | WKVIEEKAFSPEVIPMF  | A1 | FSPEVIPMF | Cw |
|      | WKVIEEKAFSPEVIPMF  | A1 | AFSPFEVIP | Cw |
|      | WKVIEEKAFSPEVIPMF  | A1 | KAFSPEVIP | Cw |
|      | TVYYGVPPWKDAETTLF  | A1 | TVYYGVPPW | A* |
|      | TVYYGVPPWKDAETTLF  | A1 | TVYYGVPPW | Cw |
|      | TVYYGVPPWKDAETTLF  | A1 | WKDAETTLF | B' |
|      | TVYYGVPPWKDAETTLF  | A1 | VWKDAETTL | A' |
|      | TVYYGVPPWKDAETTLF  | A1 | VWKDAETTL | Cw |
|      | SVKLWYQLEKEPIIGA   | A1 | SVKLWYQLE | A* |
|      | SVKLWYQLEKEPIIGA   | A1 | SVKLWYQLE | B' |
|      | LFCASDAKAYETEMHNWV | A1 | SDAKAYETEM | B' |
| A*02:01 | PPLVKLWYQLEKEPIIGA | D  | LVKLWYQLE | A* |
| A*34:02 | PPLVKLWYQLEKEPIIGA | D  | KLWYQLEKEPIIGA | A* |
| B*45:01 | PPLVKLWYQLEKEPIIGA | D | QLEKEPIIG | B* |
| B*47:01 | PPLVKLWYQLEKEPIIGA | D | YQLEKEPII | A* |
| Cw*06:02 | PPLVKLWYQLEKEPIIGA | D | YQLEKEPII | B |
| 74 | KWPKMIGGGFIKVR | D | MIGGGFIK | A* |
| KWPKMIGGGFIKVR | D | KMIIGGGGF | A* |
| KWPKMIGGGFIKVR | D | KMIIGGGGF | B |
| 75 | VIWGTKPFRFLPQIKETW | D | IVWGTKPK | A* |
| VIWGTKPFRFLPQIKETW | D | KPTFKRLPI | Cw |
| VIWGTKPFRFLPQIKETW | D | VIWGTKPKF | A* |
| 76 | RQANFLGKIWPSPHKGR | D | RQANFLGKI | B* |
| RQANFLGKIWPSPHKGR | D | RQANFLGKI | Cw |
| RQANFLGKIWPSPHKGR | D | FLGKIWPSP | A* |
| 77 | KIEELREHLRRGWGFTTPDK | D | REHLLRWGF | B* |
| KIEELREHLRRGWGFTTPDK | D | LREHLLRWG | Cw |
| KIEELREHLRRGWGFTTPDK | D | HLLRWGFTT | A* |
| 78 | GFAILKCKDKEFNGTGPCK | A1 | KEFNGTGP | B* |
| 79 | AILNIPTIRQGLERAL | D | IRQGLERAL | Cw |
| AILNIPTIRQGLERAL | D | AILNIPTRI | A* |
| AILNIPTIRQGLERAL | D | QGLERAL | B |
| 80 | QKTELQAINLALQDSGLEV | D | LALQDSGLE | A* |
| QKTELQAINLALQDSGLEV | D | TELQAINLA | B* |
| QKTELQAINLALQDSGLEV | D | QKTELQAIN | B* |
| QKTELQAINLALQDSGLEV | D | NLALQDSGL | A* |
| 81 | IIGRNLTIQIGCTLNFP | D | IGCTLNFP | A* |
| IIGRNLTIQIGCTLNFP | D | NLLTQIGCTLNFP | A* |
| IIGRNLTIQIGCTLNFP | D | LTTIQIGCTL | A* |
| IIGRNLTIQIGCTLNFP | D | TQIGCTLN | Cw |
| IIGRNLTIQIGCTLNFP | D | LTQIGCTL | Cw |
| 82 | KWPKMIGGGFIKVR | D | KMIIGGGGF | A* |
| 83 | LWQRPLVTIKIGQQLKEA | D | LWQRPLVTI | A* |
| LWQRPLVTIKIGQQLKEA | D | QRPLVTIK | Cw |
| 84 | LKEALLDTGADDTVLEEI | D | LKEALLDTG | B* |
| 85 | KRQEILDWVYHTQGYF | A1 | QEILDWV | B* |
| KRQEILDWVYHTQGYF | A1 | QEILDWV | B |
| KRQEILDWVYHTQGYF | A1 | RQEILDWV | Cw |
| KRQEILDWVYHTQGYF | A1 | ILDLWVYHT | A* |
| IYSLIEESNQOQEKNEQEL | D | | |
| 86 | SFNCGGEFFYCNTSGLF | A1 | SFNCGGEFFY | A* |
| SFNCGGEFFYCNTSGLF | A1 | SFNCGGEFF | Cw |
| SFNCGGEFFYCNTSGLF | A1 | SFNCGGEFF | C* |
| SFNCGGEFFYCNTSGLF | A1 | GEFFYCNTS | B* |
| 87 | MEKEGKISKIGPENPY | A1 | KEGKISKIGPENPY | B* |
| MEKEGKISKIGPENPY | A1 | ISKIGPEN | A* |
| 88 | ARKNRRRRWRARQROI | A1 | RRWRARQRO | Cw |
### Table 3. Peptides not predicted

| Participant’s Identification | Participant’s HLA Alleles | Experimental Peptide Sequence |
|-----------------------------|---------------------------|-------------------------------|
| E92                         | A*02:01                   | FKGPRIIKCFNGKKEGHI            |
|                             | A*30:02                   |                               |
|                             | B*44:03                   |                               |
|                             | B*14:02                   |                               |
|                             | Cw*04:01                  |                               |
|                             | Cw*08:02                  |                               |
| E95                         | A*23:01                   | LVQANPDCKSILRAL (both tin     |
|                             | A*74:01                   | SKQKTQAAADTGNSSKV             |
|                             | B*44:03                   |                               |
|                             | B*15:10                   |                               |
|                             | Cw*04:01                  |                               |
|                             | Cw*16:01                  |                               |
| L913                        | A*02:01                   | IYSLIEESOQNQEQKNQEL           |
|                             | A*34:02                   |                               |
|                             | B*45:01                   |                               |
|                             | B*47:01                   |                               |
|                             | Cw*06:02                  |                               |
|                             | Cw*16:01                  |                               |

### Table 4. Participant characteristics, HIV-1 infecting stage clade and HLA class I haplotypes.

| Subject | Sex | Age range (years) | HIV-1 subtype | Class-I HLA | Early Time Point (Days) | Fiebig Staging | Lal Poir |
|---------|-----|-------------------|---------------|-------------|-------------------------|----------------|----------|
| 91      | M   | 31-40             | A             | A*0201,*0301;B*5301,*5802;Cw*0401,*060 | 121           | VI       |
| 92      | F   | 21-30             | D             | A*0201,*3002;B*4403,*1402;Cw*0401,*0802 | 52            | VI       |
| 94      | M   | 51-60             | A             | A*3402,*7401;B*4403,*5802;Cw*0401,*060 | 28            | V        |
| 95      | M   | 21-30             | A             | A*2301,*7401;B*4403,*1510;Cw*0401,*160 | 30            | VI       |
| 913     | F   | 11-20             | D             | A*0201,*3402;B*4501,*4701;Cw*0602,*160 | 61            | VI       |
| 914     | F   | 21-30             | D             | A*0101,*0201;B*0702,*4415;Cw*0407,*070 | 31            | IV       |

**Figures**
Figure 1

NetMHCpan Binder Predictions. (A) Using our experimental peptide sequences as inputs into NetMHCpan4.0 to predict epitopes for 22 HLA types represented in the 6 HIV-1 Infected people, a heatmap showing absolute counts of computationally predicted 9-mer binders against HIV-1 genes was constructed. The dendrogram shows the nearest similarity for the number of predicted counts across HLA types; (B) the length of the HIV-1 protein sequence plotted against the absolute number of NetMHCpan4.0 predicted 9mer binders showing a positive correlation (Spearman’s correlation coefficient, rs=0.88). The number distinct predictions dependent on the length of the HIV-1 sequence ; (C) comparison of HIV-1 clade A and D absolute number of NetMHCpan4.0 predicted 9mer binders per HIV-1 gene for the wet experiment test peptide sequences. The algorithm predicted more
binders for clade D than clade A.

Figure 2
Computational epitope prediction. NetMHCpan4.0 set length plotted against the number of predicted binders per HLA type shows that the number of predictions reduces as the input set length increases. The dotted line is the trend line, whereas the solid line is the line of best fit. The core 9mer epitope sequence anchoring at positions 2 and 9 was similar across the 9mer through 14mer set length except for one 14-mer peptide (hit 72 in table 2).
Figure 3

Early versus Late Peptides. Experimentally mapped peptides at baseline (n=34) and at least 12 months later (n=34) were compared using the 9-mer computational NetMHCpan4.0 scores of the hits. The lower the computational score the stronger the predicted binding. Late peptides were significantly stronger binders than early peptides (Wilcoxon signed rank test, \( p=0.0000005 \)).
Figure 4

ROC plot. False versus true positive rate for all 9-mer and a single 14-mer test peptides across the 22 test HLA class I types. The diagonal line shows the random guess whereas the red curve shows the observed experimentally mapped epitopes versus the NetMHCpan4.0 expected predictions.
ELISPOT peptide consort; the experimental peptide mapping data was generated from culture ELISPOT of multiple peptides pools were tested in duplicate wells per time point followed by ex-vivo ELISPOT of potential candidate epitopes. To experimentally map a single time point required at least 400 assay wells.