

## RESEARCH ARTICLE

## Open Access



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

Daniel Lule Bugembe<sup>1\*</sup> , Andrew Obuku Ekii<sup>1</sup>, Nicaise Ndembu<sup>2</sup>, Jennifer Serwanga<sup>1,3</sup>, Pontiano Kaleebu<sup>1,3</sup> and Pietro Pala<sup>1</sup>

## 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- $\gamma$  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

\* Correspondence: [dan.lule@mrcuganda.org](mailto:dan.lule@mrcuganda.org)

<sup>1</sup>MRC/UVRl and LSHTM Uganda Research Unit, P. O. Box 49, Plot 51-59 Nakiwogo Road, Entebbe, Uganda

Full list of author information is available at the end of the article



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 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 1) at 2 time points each was used for comparison with the 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 1), enrolled under the following criteria: (i) detection of HIV-1 P24 antigen with a simultaneous negative HIV-1 antibody

**Table 1** Participant characteristics, HIV-1 infecting clade, Fiebig stage and HLA class I haplotypes

Subject	Sex	Age range (years)	HIV-1 subtype	Class-I HLA	Early Time Point (Days)	Fiebig Staging	Late Time Point (Days)
91	M	31–40	A	A*0201,*0301;B*5301,*5802;Cw*0401,*0602	121	VI	841
92	F	21–30	D	A*0201,*3002;B*4403,*1402;Cw*0401,*0802	52	VI	743
94	M	51–60	A	A*3402,*7401;B*4403,*5802;Cw*0401,*0602	28	V	358
95	M	21–30	A	A*2301,*7401;B*4403,*1510;Cw*0401,*1601	30	VI	570
913	F	11–20	D	A*0201,*3402;B*4501,*4701;Cw*0602,*1601	61	VI	211
914	F	21–30	D	A*0101,*0201;B*0702,*4415;Cw*0407,*0702	31	IV	181

ELISA (2 participants) or documented HIV-1 seronegative 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 (Fig. 1), were 17aa long, overlapping by 11aa and spanning the HIV-1 proteome consensus for subtypes A1 and D. Cultured ELISPOT assays using 200,000 cells/well as previously documented by Obuku AE. et.al [34]. and ex-vivo IFN-γ ELISPOT assay using 100,000 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.

**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].

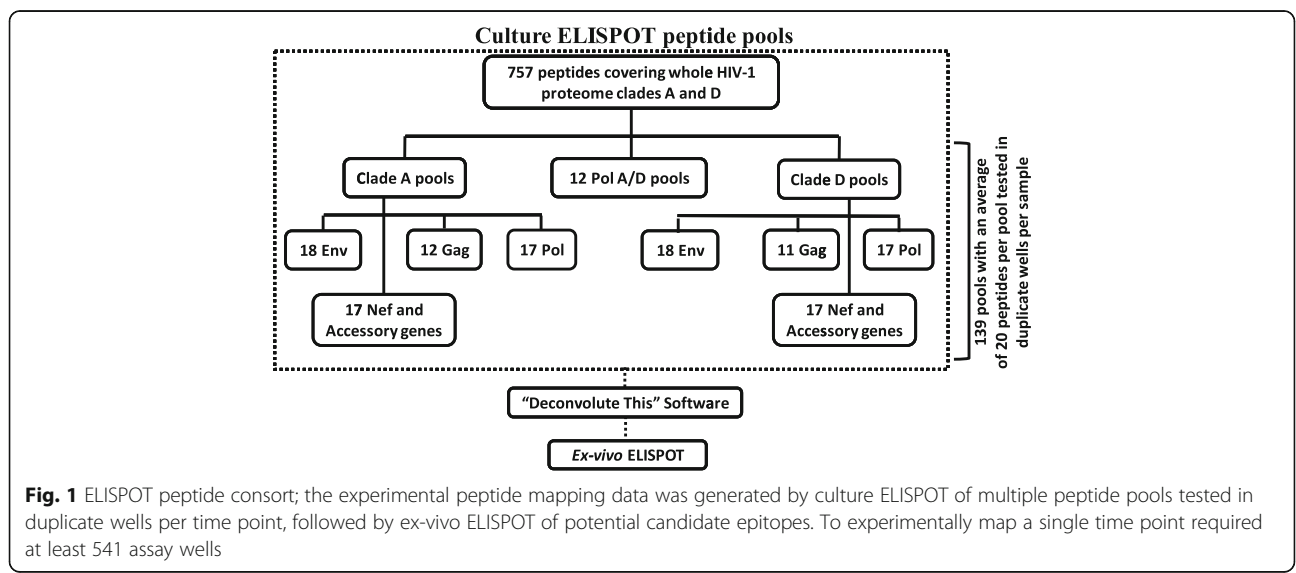
**Computation epitope prediction**

HIV-1 subtypes A1 and D 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](http://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 1) 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 were 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**

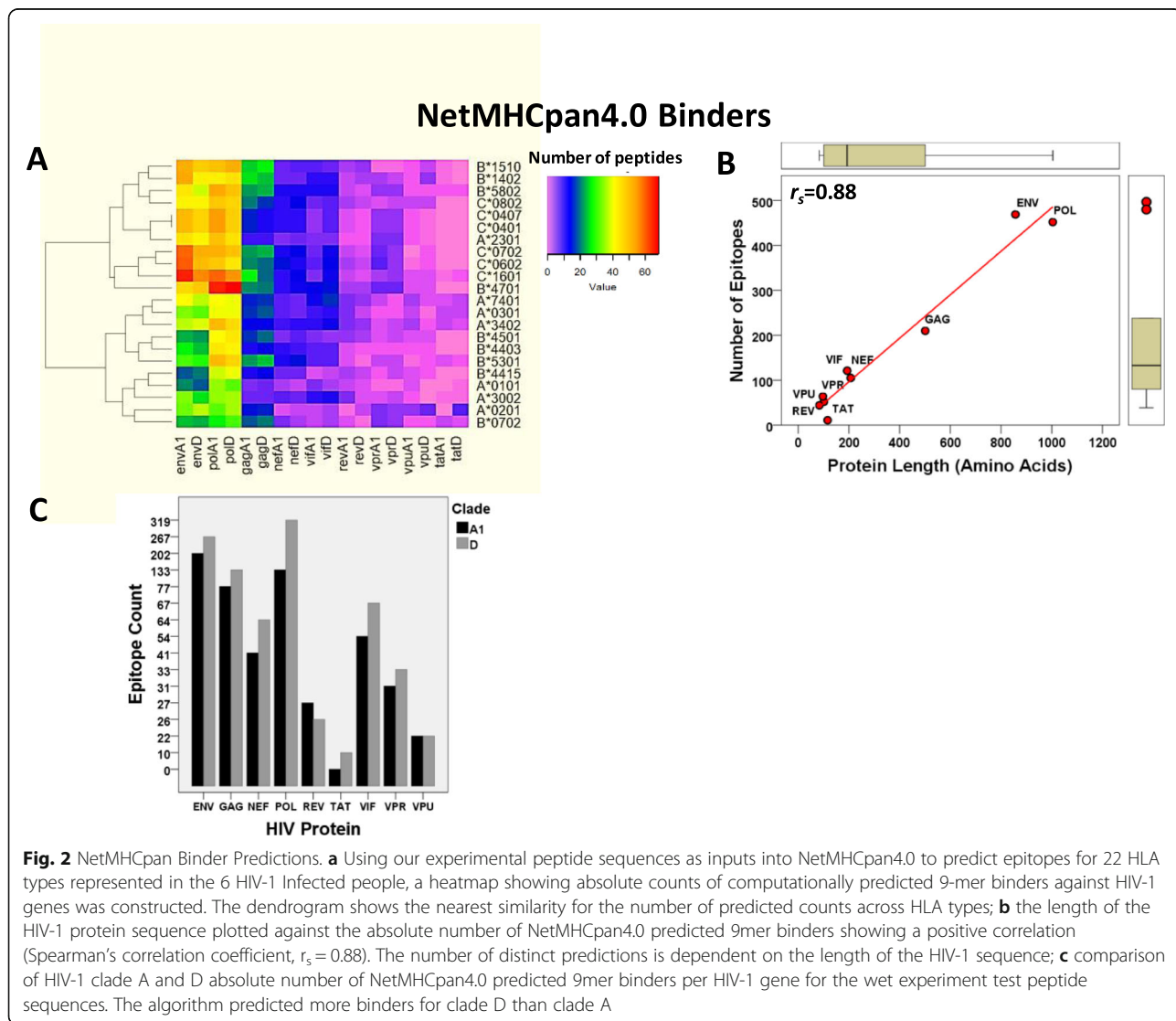
Statistics computations and plots were generated using SPSS version 24.0.0.0. The NetMHCpan4.0 computational performance was evaluated using a confusion



**Fig. 1** ELISPOT peptide consort; the experimental peptide mapping data was generated by culture ELISPOT of multiple peptide pools 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 541 assay wells

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 A1 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 HIV-1 clades A1 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 consensus 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



**Fig. 2** 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 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,  $r_s = 0.88$ ). The number of distinct predictions is 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



**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0% Rank	MHCflurry 9-mer Epitope prediction	MHCflurry 9-mer HLA prediction	MHCflurry 9-mer prediction (µm)	NetCTL 9-mer Epitope prediction
	B*44:03	14	HQDPIPKQSSQPRGD	D	HQDPIPKQP	Cw*04:01	0.60	HQDPIPKQP	Cw*04:01	7.89	
	B*58:02							LEAQEEEEV	Cw*06:02	8.72	
	Cw*04:01										
	Cw*06:02										
E95	A*23:01	15	VAVHVASGYIAEVIPA	A1	VAVHVASGY	Cw*16:01	1.50	VAVHVASGY	A*23:01	30.27	
	A*74:01	17	KRWIILGLNKIVRMYSVP	A1	WIILGLNKI	A*23:01	0.60	WIILGLNKI	A*23:01	6.99	
	B*44:03	16	KRWIILGLNKIVRMYSVP	A1	IILGLNKIV	A*74:01	0.60	IILGLNKIV	B*44:03	28.76	
	B*15:10	17	NMMLNIVGGHQAAAMQMLK	A1	HQAAAMQMLK	B*15:10	0.17				
	Cw*04:01				HQAAAMQMLK	A*74:01	0.90				HQAAMQMLK
	Cw*16:01				HQAAAMQMLK	Cw*04:01	1.30	HQAAMQMLK	Cw*04:01	26.63	
		18	KNWMTETLLVQNPDPCK	A1	TETLLVQNA	B*44:15	0.09	TETLLVQNA	B*44:03	5.85	
					KNWMTETLLVQNPDPCK	A*23:01	0.80	KNWMTETLL	A*23:01	25.75	
		19	FRDYDRFFKTLRAEQA	A1	FRDYDRFF	Cw*04:01	0.03	FRDYDRFF	Cw*04:01	4.56	
					FRDYDRFFKTLRAEQA	A*23:01	0.60	FRDYDRFF	A*23:01	4.56	
					FRDYDRFFKTLRAEQA	Cw*04:01	1.10	FRDYDRFF	Cw*04:01	0.68	
		20	GATLEEMMTACQGVGGPGH	A1	EEMMTACQG	B*44:03	0.25	EEMMTACQG	B*44:03	0.97	
		21	LRALGPGATLEEMMTA	A1	RALGPGATL	B*15:10	1.80	RALGPGATL	B*44:01	0.56	
					LRALGPGATLEEMMTA	Cw*04:01	0.60	RALGPGATL	Cw*04:01	0.56	
		22	FFKTLRAEQATQEVKNWMM	A1	AEQATQEVK	B*44:03	0.15	AEQATQEVK	B*44:03	8.59	
		23	MEKEGKISKIPENPY	A1	SKIGPENPY	B*15:03	0.50	SKIGPENPY	A*23:01	35.68	
					MEKEGKISKIPENPY	B*15:10	0.50	SKIGPENPY	B*44:03	9.49	
		25	WKVIEEKAFSPEVPMF	A1	AFSPEVPMF	A*23:01	0.40	AFSPEVPMF	A*23:01	5.95	
					WKVIEEKAFSPEVPMF	A*23:01	1.70	WKVIEEKA	A*23:01	17.53	
					WKVIEEKAFSPEVPMF	B*44:03	0.80	EKAFSPEV	B*44:03	3.38	EKAFSPEV
					WKVIEEKAFSPEVPMF	B*44:15	0.03	EKAFSPEV			FSPEVPMF
					WKVIEEKAFSPEVPMF	Cw*04:01	0.50	FSPEVPMF	Cw*04:01	1.63	FSPEVPMF
					WKVIEEKAFSPEVPMF	Cw*16:01	0.50	FSPEVPMF	A*23:01	0.46	
					WKVIEEKAFSPEVPMF	Cw*16:01	1.20	KAFSPEV	B*44:03	35.68	
		25	HQMKDCTERQANFLGKIW	A1	ROANFLGKI	B*44:03	1.00	ROANFLGKI	B*44:03	11.67	
		26	PMFSALSEGATPQDLNMM	A1	SEGATPQDL	B*44:03	0.80	SEGATPQDL	B*44:03	0.33	
		27	HLARNCRAPRKGCKWK	A1	HLARNCRAP	A*74:01	0.60	HLARNCRAP	A*23:01	35.46	

**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0% Rank	MHCflurry 9-mer Epitope prediction	MHCflurry 9-mer HLA prediction	MHCflurry affinity ( $\mu$ m)	NetCTL 9-mer Epitope prediction
			LYQMANPDCSKILRAL	A1							
E913	A*02:01	28	VATLYCVHQRIDVKDTK	A1	ATLYCVHQR	A*74:01	0.90	ATLYCVHQR	A*23:01	21.03	
	A*34:02	29	KIEEQNKSKOKTQOAAA	A1	EIQNKSKQK	A*74:01	1.03	EIQNKSKQK	C*04:01	10.98	
	B*45:01	30	AGPIPPGQMREPRGSDIA	A1	AGPIPPGQM	B*15:10	0.60	AGPIPPGQM	C*04:01	2.02	
	B*47:01		SKQKTQAAADTGNSSKV	A1							
	Cw*06:02	31	LWQRPLVTIKIGGQLKEA	D	LWQRPLVTI	A*02:01	1.60	LWQRPLVTI	A*02:01	11.25	
	Cw*16:01		LWQRPLVTIKIGGQLKEA	D	QRPLVTIKI	Cw*06:02	0.70	QRPLVTIKI	Cw*06:02	16.29	
E914	A*01:01	33	DKWASLWNWFISITQWLWY	D	FSITQWLWY	A*01:01	0.06	FSITQWLWY	B*07:02	24.05	FSITQWLWY
	A*02:01		DKWASLWNWFISITQWLWY	D	KWASLWNWF	Cw*04:07	1.20				
	B*07:02		DKWASLWNWFISITQWLWY	D	SLWNWFSIT	A*02:01	1.66				
	B*44:15	34	PVDPDEVEKATEGENNSL	A1	ATEGENNSL	A*01:01	1.74				
	Cw*04:07										
	Cw*07:02										
L91	A*02:01	35	EQMHTDIISLWDQSLK	A1	IISLWDQSLK	A*03:01	1.90	IISLWDQSLK	A*03:01	20.55	ISLWDQSLK
	A*03:01		EQMHTDIISLWDQSLK	A1	MHTDIISLW	B*58:02	0.90	MHTDIISLW	A*03:01	23.95	
	B*53:01		EQMHTDIISLWDQSLK	A1	MHTDIISLW	Cw*06:02	1.30	MHTDIISLW	Cw*06:02	3.72	
	B*58:02		EQMHTDIISLWDQSLK	A1	QMHTDIISL	A*02:01	1.90	QMHTDIISL	B*53:01	18.52	
	Cw*04:01		EQMHTDIISLWDQSLK	A1	QMHTDIISL	B*53:01	1.20	QMHTDIISL	A*03:01	0.06	KIRLRPGGK
	Cw*06:02	36	LETSEGCKQIGQLQPAI	D	ILAQLQPAI	A*02:01	0.40	IMEQLQPAL	A*03:01	21.06	
		37	SGGKLDAWEKIRLRPGGK	A1	KIRLRPGGK	A*03:01	0.25	KIRLRPGGK	A*03:01	0.19	KIRLRPGGK
		38	LETTEGCQIMEQLQPAL	A1	IMEQLQPAL	A*03:01	0.40	IMEQLQPAL	A*03:01	0.11	
			LETTEGCQIMEQLQPAL	A1	IMEQLQPAL	Cw*04:01	0.80	IMEQLQPAL	Cw*04:01		
			LETTEGCQIMEQLQPAL	A1	QIMEQLQPA	A*02:01	0.70				
		39	ERILSTCLGRSAEPVPL	A1	RSAPVPL	B*58:02	0.12	RILSTCLGR	A*03:01	0.11	
			ERILSTCLGRSAEPVPL	A1	RILSTCLGR	A*03:01	0.90	RILSTCLGR	A*03:01		
			ERILSTCLGRSAEPVPL	A1	CLGRSAEPV	A*02:01	2.00				
		40	LVGTPVNIIGNRMILTQI	A1	LVGTPVNI	A*02:01	1.61	LVGTPVNI	:01	4.69	

**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0% Rank	MHCflurry 9-mer Epitope prediction	MHCflurry HLA prediction	MHCflurry 9-mer prediction	MHCflurry affinity ( $\mu$ m)	NetCTL 9-mer Epitope prediction	
L92	A*02:01 A*30:02 B*44:03 B*14:02 Cw*04:01 Cw*08:02	41	CKQIQQLQPAIQTGSEEL	D	QIQQLQPA	A*02:01	1.80	AIQTGSEEL	A*03:01	A*03:01	21.04		
		42	PAIQTGSEELRSLYNTVA	D	AIQTGSEEL	A*03:01	1.50	AIQTGSEEL	Cw*06:02	Cw*06:02	3.74		
			PAIQTGSEELRSLYNTVA	D	AIQTGSEEL	A*03:01	1.10	AIQTGSEEL	Cw*06:02	Cw*06:02	5.48		
			PAIQTGSEELRSLYNTVA	D	AIQTGSEEL	A*03:01	1.50	AIQTGSEEL	Cw*06:02	Cw*06:02	5.48		
			NDIQKLVGKLNWASQIYP	D	LRSLYNTVA	Cw*06:02	0.70	LRSLYNTVA	KLNWASQIY	A*30:02	A*30:02	1.22	KLNWASQIY
			NDIQKLVGKLNWASQIYP	D	LRSLYNTVA	Cw*06:02	0.50	LRSLYNTVA	KLNWASQIY	A*02:01	A*02:01	11.10	
			NDIQKLVGKLNWASQIYP	D	LRSLYNTVA	Cw*06:02	0.90	LRSLYNTVA	KLNWASQIY	A*02:01	A*02:01	13.24	
			NDIQKLVGKLNWASQIYP	D	LRSLYNTVA	Cw*06:02	1.80	LRSLYNTVA	KLNWASQIY	A*02:01	A*02:01	0.95	
			LWKTYWGLHTGEREWHL	D	LWKTYWGL	A*02:01	1.70	LWKTYWGL	YSPLSFQTL	Cw*04:01	Cw*04:01	2.47	RQGYSPLSF
			LWKTYWGLHTGEREWHL	D	LWKTYWGL	A*02:01	1.50	LWKTYWGL	YSPLSFQTL	Cw*04:01	Cw*04:01	4.51	
L94	A*34:02 A*74:01 B*44:03 B*58:02 Cw*04:01 Cw*06:02	45	SLVNRVRQGYSPLSFQTL	D	NRVRQGYSP	B*14:02	0.12	RQGYSPLSF	A*30:02	A*30:02	3.12	CRIKQIINM	
			SLVNRVRQGYSPLSFQTL	D	NRVRQGYSP	B*14:02	0.70	RQGYSPLSF	A*30:02	A*30:02	28.36		
			SLVNRVRQGYSPLSFQTL	D	NRVRQGYSP	B*14:02	1.20	RQGYSPLSF	A*30:02	A*30:02	3.17		
			SLVNRVRQGYSPLSFQTL	D	NRVRQGYSP	B*14:02	1.40	RQGYSPLSF	A*30:02	A*30:02	32.11		
			TLPCRKQIINMWQGV	D	CRIKQIINM	A*02:02	0.40	CRIKQIINM	EMKNCFSNI	B*44:03	B*44:03	7.71	
			MRVRIQRYQHLLWRW	D	RNYQHLLWRW	B*44:03	0.40	RNYQHLLWRW	EMKNCFSNI	B*44:03	B*44:03	4.68	
			GEMKNCFSNIITEIRDKK	D	EMKNCFSNI	B*44:03	0.30	EMKNCFSNI	YFDCFSDSA	Cw*04:01	Cw*04:01	2.52	
			NVTENFNWKNMWVEQMH	D	NFNWKNNM	Cw*04:01	1.06	NFNWKNNM	YFDCFSDSA	Cw*06:02	Cw*06:02	14.91	
			NVTENFNWKNMWVEQMH	D	NFNWKNNM	Cw*04:01	1.81	NFNWKNNM	YFDCFSDSA	Cw*06:02	Cw*06:02	0.78	
			WLIDRERAEEDSGNESE	D	TENFNWKNNM	B*44:03	2.00	TENFNWKNNM	YFDCFSDSA	Cw*06:02	Cw*06:02	7.89	
L95	A*23:01	50	WLIDRERAEEDSGNESE	D	WLIDRER	A*02:01	2.00	WLIDRER	A*02:01	A*02:01	4.68		
		51	LIHLHYFDFCFSDSAIRKA	A1	YFDCFSDSA	Cw*04:01	0.90	YFDCFSDSA	Cw*04:01	Cw*04:01	2.52		
			LIHLHYFDFCFSDSAIRKA	A1	YFDCFSDSA	Cw*04:01	1.60	YFDCFSDSA	Cw*06:02	Cw*06:02	14.91		
			LIHLHYFDFCFSDSAIRKA	A1	YFDCFSDSA	Cw*06:02	1.40	YFDCFSDSA	Cw*06:02	Cw*06:02	0.78		
			LIHLHYFDFCFSDSAIRKA	A1	YFDCFSDSA	Cw*04:01	1.10	YFDCFSDSA	Cw*04:01	Cw*04:01	0.78		
			LIHLHYFDFCFSDSAIRKA	A1	YFDCFSDSA	Cw*04:01	1.10	YFDCFSDSA	Cw*04:01	Cw*04:01	0.78		
			HLARNCRAPRKGCGWK	A1	ARNCRAPRK	A*34:02	1.10	ARNCRAPRK	FSDSAIRKA	Cw*04:01	Cw*04:01	0.78	
			HLARNCRAPRKGCGWK	A1	ARNCRAPRK	A*34:02	1.50	ARNCRAPRK	FSDSAIRKA	Cw*04:01	Cw*04:01	0.78	
			HLARNCRAPRKGCGWK	A1	ARNCRAPRK	A*34:02	0.60	ARNCRAPRK	FSDSAIRKA	Cw*04:01	Cw*04:01	0.78	
			HLARNCRAPRKGCGWK	A1	ARNCRAPRK	A*34:02	0.60	ARNCRAPRK	FSDSAIRKA	Cw*04:01	Cw*04:01	0.78	
53	SKQKTOAAAADTGNSSKV	A1	AADTGNSSK	A*34:02	1.15	AADTGNSSK	HQDPIKQOP	Cw*04:01	Cw*04:01	7.89			
54	HQDPIKQOPSSOPRGD	D	HQDPIKQOP	Cw*04:01	0.60	HQDPIKQOP	WLILGLNKI	A*23:01	A*23:01	6.99			
55	KRWILGLNKIVRMYSVP	A1	WLILGLNKI	A*23:01	0.60	WLILGLNKI							



**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0% Rank	MHCflurry 9-mer Epitope prediction	MHCflurry 9-mer HLA prediction	MHCflurry 9-mer prediction	MHCflurry affinity ( $\mu$ m)	NetCTL 9-mer Epitope prediction
A*74:01			KRWILGLNKIVRMYSPV	A1	IILGLNKIV	A*74:01	0.60	IILGLNKIV	Cw*04:01		2.81	
B*44:03	56		NMMLNIVGGHQAAAMQMLK	A1	GHQAAAMQML	B*15:10	0.40					
B*15:10			NMMLNIVGGHQAAAMQMLK	A1	HQAAAMQMLK	A*74:01	0.90					HOAAAMQMLK
Cw*04:01			NMMLNIVGGHQAAAMQMLK	A1	HQAAAMQMLK	Cw*04:01	1.30	HQAAAMQMLK	Cw*04:01		26.63	
Cw*16:01	57		EVNIVTDSQYALGIQA	A1	EVNIVTDSQ	B*44:03	0.50	EVNIVTDSQ	B*44:03		19.78	
	58		AYETEMHNWVWATHACV	A1	TEMHNWVAT	B*44:03	0.40	TEMHNWVAT	B*44:03		0.88	YETEMHNWV
			AYETEMHNWVWATHACV	A1	MHNWVATHA	B*15:10	0.90	MHNWVATHA	Cw*04:01		11.32	
			AYETEMHNWVWATHACV	A1	MHNWVATHA	B*44:03	0.90	MHNWVATHA	B*44:04		22.06	
	59		AAEWDRLHPVHAGPI	A1	AAEWDRLHP	B*44:03	0.50	AAEWDRLHP	B*44:03		34.74	AEWDRLHPV
			AAEWDRLHPVHAGPI	A1	LHPVHAGPI	B*15:10	0.15	AAEWDRLHP	A*23:01		34.79	
			AAEWDRLHPVHAGPI	A1	RLHPVHAGP	A*74:01	1.50	AAEWDRLHP	Cw*04:01		6.59	
	60		LRALGPGATLEEMMTA	A1	RALGPGATL	B*15:10	0.60	RALGPGATL	A*23:01		8.98	
			LRALGPGATLEEMMTA	A1	RALGPGATL	Cw*04:01	0.60	RALGPGATL	Cw*04:01		0.56	
	61		FFKTLRAEQATQEVKNWM	A1	AEQATQEVK	B*44:03	0.15	AEQATQEVK	B*44:03		8.59	
	62		GTTSTPQEQIGWMTGNPPI	A1	QEQIGWMTG	B*44:15	0.65	QEQIGWMTG	B*44:03		1.85	
			GTTSTPQEQIGWMTGNPPI	A1	GWMTGNPPI	A*23:01	1.40	GWMTGNPPI	A*23:01		0.75	
	63		WVKVIEKAFSPEVPMF	A1	EKAFSPEV	B*44:15	0.62	EKAFSPEV				EKAFSPEV
			WVKVIEKAFSPEVPMF	A1	EKAFSPEV	B*44:03	0.80	EKAFSPEV	B*44:03		3.38	
			WVKVIEKAFSPEVPMF	A1	WVKVIEEKA	A*23:01	1.70	WVKVIEEKA	A*23:01		34.76	
			WVKVIEKAFSPEVPMF	A1	FSPEVPMF	Cw*04:01	0.50	FSPEVPMF	Cw*04:01		1.63	
			WVKVIEKAFSPEVPMF	A1	AFSPEVPM	Cw*16:01	0.70	AFSPEVPM	Cw*04:01		0.28	
			WVKVIEKAFSPEVPMF	A1	KAFSPEVIP	Cw*16:01	1.20	KAFSPEVIP	Cw*04:01		7.47	
	64		TVYGVVWVKDAETTLF	A1	TVYGVVW	A*74:01	0.17	TVYGVVW	B*44:03		7.96	VYGVVWVK
			TVYGVVWVKDAETTLF	A1	TVYGVVW	Cw*16:01	1.10	TVYGVVW	Cw*04:01		1.25	
			TVYGVVWVKDAETTLF	A1	WKDAETTLF	B*15:10	0.90	WKDAETTLF	A*23:01		4.36	
			TVYGVVWVKDAETTLF	A1	WKDAETTL	A*23:01	0.60	WKDAETTLF	A*23:01		1.97	
			TVYGVVWVKDAETTLF	A1	WKDAETTL	Cw*04:01	0.60	WKDAETTL	Cw*04:01		3.97	
	65		LRWGTMLGMIIICSA	A1	RWGTMLGM	A*23:01	0.90	RWGTMLGM	A*23:01		1.34	
			LRWGTMLGMIIICSA	A1	RWGTMLGM	Cw*04:01	0.94	RWGTMLGM	Cw*04:01		32.48	
	66		GHQAAAMQMLKDTINEEAA	A1	HQAAAMQMLK	A*74:01	0.90	HQAAAMQMLK	A*23:01		10.14	
	67		IKQGPKEPRDYDRFFK	A1	FRDYDRFF	A*23:01	0.60	FRDYDRFF	A*23:01			

**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0 Rank	MHCflurry 9-mer Epitope prediction	MHCflurry 9-mer HLA prediction	MHCflurry 9-mer affinity (µm)	NetCTL 9-mer Epitope prediction
			IKQKPEFRDYVDRFK	A1	FRDYVDRFF	Cw*04:01	0.60	FRDYVDRFF	Cw*04:01	0.68	
		68	FRDYVDRFKTLRAEQA	A1	FRDYVDRFF	A*23:01	0.60	FRDYVDRFF	A*23:01	10.14	
			LVDNAMPDCKSLRAL	A1							
		69	MREPRGSDIAGTTSTPQEIQI	A1	MREPRGSDI	B*15:10	2.00	MREPRGSDI	Cw*04:01	1.95	
		70	EKIRLRPGGKKYRLKHL	A1	RLRPGKKK	A*74:01	0.28	RLRPGKKK	A*23:01	28.99	
		71	VATLYCVHQRIDVKDTK	A1	ATLYCVHQR	A*74:01	0.90	ATLYCVHQR	B*44:03	28.12	
		72	LFCASDAKAYETEMHNW	A1	SDAKAYETEMHNW	B*44:03	0.31	KAYETEMHN	B*44:03	35.92	
L913	A*02:01	73	PPLVKLWYQLEKEPIGA	D	LVKLWYQLE	A*34:01	0.50	LVKLWYQLE	A*02:01	18.93	
	A*34:02		PPLVKLWYQLEKEPIGA	D	KLWYQLEKEPIGA	A*02:01	1.47	WYQLEKEPI	A*02:01	18.24	
	B*45:01		PPLVKLWYQLEKEPIGA	D	QLEKEPIIG	B*45:01	0.90	QLEKEPIIG	B*45:01	17.25	
	B*47:01		PPLVKLWYQLEKEPIGA	D	YQLEKEPII	A*02:01	0.80	YQLEKEPII	A*02:01	0.25	
	Cw*06:02		PPLVKLWYQLEKEPIGA	D	YQLEKEPII	B*47:01	1.20	YQLEKEPII	Cw*06:02	2.05	
	Cw*16:01	74	KWKPKMIGGIGGFIKVR	D	MIGGIGGFIK	A*34:02	0.20	MIGGIGGFIK	A*02:01	1.55	
			KWKPKMIGGIGGFIKVR	D	KMIGGIGGF	A*02:01	1.00	KMIGGIGGF	A*02:01	5.37	
			KWKPKMIGGIGGFIKVR	D	KMIGGIGGF	B*47:01	1.10	KMIGGIGGF	Cw*06:02		
		75	VWVGKTPKFRLPQKETW	D	IVWGTPK	A*34:02	0.15	KTPKFRLP	A*02:01	4.43	
			VWVGKTPKFRLPQKETW	D	KTPKFRLP	Cw*16:01	1.10	KTPKFRLP	A*02:01	7.82	
			VWVGKTPKFRLPQKETW	D	VWVGKTPKF	A*34:02	1.00	VWVGKTPKF	Cw*06:02		
		76	ROANFLGKIWPSHKGR	D	ROANFLGKI	B*47:01	0.40	ROANFLGKI	Cw*06:02	9.23	
			ROANFLGKIWPSHKGR	D	ROANFLGKI	Cw*06:02	2.00	ROANFLGKI	B*45:01	7.03	
			ROANFLGKIWPSHKGR	D	FLGKIWPSH	A*34:02	1.10	REHLLRWGF	A*02:01	20.74	
		77	KIEELREHLLRWGFTTPDK	D	REHLLRWGF	B*47:01	0.03	REHLLRWGF	B*45:01	1.11	
			KIEELREHLLRWGFTTPDK	D	REHLLRWGF	B*45:01	0.70	REHLLRWGF	Cw*06:02	19.44	
			KIEELREHLLRWGFTTPDK	D	LREHLLRWG	Cw*06:02	1.40	LREHLLRWG	A*02:01	0.20	
			KIEELREHLLRWGFTTPDK	D	HLLRWGFTT	A*02:01	1.20	HLLRWGFTT	B*45:01	1.25	
		78	GFAILKCKDKFNFTGPK	A1	KEFNFTGPC	B*45:01	1.50	KEFNFTGPC	Cw*06:02	0.71	
		79	AILNIPRIROGLERALL	D	IROGLERAL	Cw*06:02	0.60	IROGLERAL	B*45:01	13.50	
			AILNIPRIROGLERALL	D	AILNIPRI	A*02:01	0.60	ROGLERALL	A*02:01	22.65	
			AILNIPRIROGLERALL	D	ROGLERALL	B*47:01	1.70	TELOAINLA	B*45:01	0.15	
		80	QKTELOAINLALQDSGLEV	D	LALQDSGLE	A*02:01	1.50	LALQDSGLE	B*45:01		
			QKTELOAINLALQDSGLEV	D	TELOAINLA	B*47:01	0.60				

**Table 2** Experimentally Mapped Peptides and Computationally Predicted Epitopes (Continued)

ID	Participant's HLA Types	Hit No	Screening Peptide	Screening Peptide HIV-1 Clade	NetMHCpan4.0 9-mer Epitope Prediction	NetMHCpan4.0 9-mer HLA Prediction	NetMHCPan4.0 Rank	MHCflurry 9-mer Epitope prediction	MHCflurry 9-mer HLA prediction	MHCflurry 9-mer prediction	MHCflurry affinity ( $\mu$ m)	NetCTL 9-mer Epitope prediction
			QKTELOAINLALQDSGLEV	D	QKTELOAIN	B*45:01	1.00	QKTELOAIN	B*45:01		12.04	
			QKTELOAINLALQDSGLEV	D	NLALQDSGL	A*34:02	1.50	NLALQDSGL	A*02:01		4.72	
		81	IIGRNLLTQIGCTLNFPPI	D	IGCTLNFPPI	A*02:01	0.90	IGCTLNFPPI	A*02:01		4.79	
			IIGRNLLTQIGCTLNFPPI	D	NLLTQIGCTLNFPPI	A*02:01	1.63					
			IIGRNLLTQIGCTLNFPPI	D	LLTQIGCTL	A*02:01	1.90	LLTQIGCTL	A*02:01		0.53	
			IIGRNLLTQIGCTLNFPPI	D	TQIGCTLNFP	Cw*16:01	1.40	TQIGCTLNFP	Cw*06:02		2.26	
			IIGRNLLTQIGCTLNFPPI	D	LLTQIGCTL	Cw*16:01	0.90					
		82	KWKPKMIGGIGGFVKR	D	KMIGGIGGF	A*02:01	1.00	KMIGGIGGF	A*02:01		1.55	
		83	LWQRPLVTIKIGGQLKEA	D	LWQRPLVTI	A*02:01	1.60	LWQRPLVTI	A*02:01		11.25	
			LWQRPLVTIKIGGQLKEA	D	QRPLVTIKI	Cw*06:02	0.70	LWQRPLVTI	Cw*06:02		0.74	
		84	LKEALLDTGADDTVLEEI	D	LKEALLDTG	B*45:01	1.20	LKEALLDTG	B*45:01		12.15	
		85	KRQEILDWVYHTQGYF	A1	QEILDWVY	B*45:01	1.70	QEILDWVY	B*45:01		1.02	
			KRQEILDWVYHTQGYF	A1	QEILDWVY	B*47:01	0.90					
			KRQEILDWVYHTQGYF	A1	QEILDWVY	Cw*06:02	1.10	QEILDWVY	Cw*06:02		21.81	
			KRQEILDWVYHTQGYF	A1	ILDLWVYHT	A*02:01	0.70	ILDLWVYHT	B*45:01		13.80	
			IYSIJEESSONQEKNEQEL	D								
L914	A*01:01	86	SFNCGGEFFYCNISGLF	A1	SFNCGGEFFY	A*01:01	0.25	SFNCGGEFFY	B*44:03		7.72	
	A*02:01		SFNCGGEFFYCNISGLF	A1	SFNCGGEFF	Cw*04:07	1.30	SFNCGGEFF	B*07:02		22.45	
	B*07:02		SFNCGGEFFYCNISGLF	A1	SFNCGGEFF	Cw*07:02	1.10	SFNCGGEFF	B*44:03		1.80	
	B*44:03		SFNCGGEFFYCNISGLF	A1	GEFFYCNTS	B*44:03	1.30	GEFFYCNTS	B*44:03		39.08	
	Cw*04:07	87	MEKEGKISKIGPENPY	A1	KEGKISKIGPENPY	B*44:03	1.30	KISKIGPEN	B*07:02		27.27	
	Cw*07:02		MEKEGKISKIGPENPY	A1	ISKIGPEN	A*01:01	1.80	ISKIGPEN	Cw*07:02		19.17	
		88	ARKNRRRRRARQRQI	A1	RRWRARQRQ	Cw*07:02	0.60	RRWRARQRQ	Cw*07:02			

Experimentally mapped peptides for all participants and their cognate computational core 9-mer and a single 14-mer epitope sequence with scores. Peptides shown in *italic* text were not algorithmically predicted as binders. Multiple computational predictions contained in a single experimental peptide were counted as a single hit. Participant's identifiers (ID) beginning with E or L represent early or late time sampling points respectively

**Table 3** Peptides not predicted

Participant's Identification	Participant's HLA Alleles	Experimental Peptide Sequence
E92	A*02:01	FKGPRKIIKCFNCGKEGHI
	A*30:02	
	B*44:03	
	B*14:02	
	Cw*04:01	
	Cw*08:02	
E95	A*23:01	LVQANPDKSILRAL (both time points)
	A*74:01	
	B*44:03	SKQKTQQAADTGNSKIV
	B*15:10	
	Cw*04:01	
	Cw*16:01	
L913	A*02:01	IYSLIEESQNQQEKNEQEL
	A*34:02	
	B*45:01	
	B*47:01	
	Cw*06:02	
	Cw*16:01	

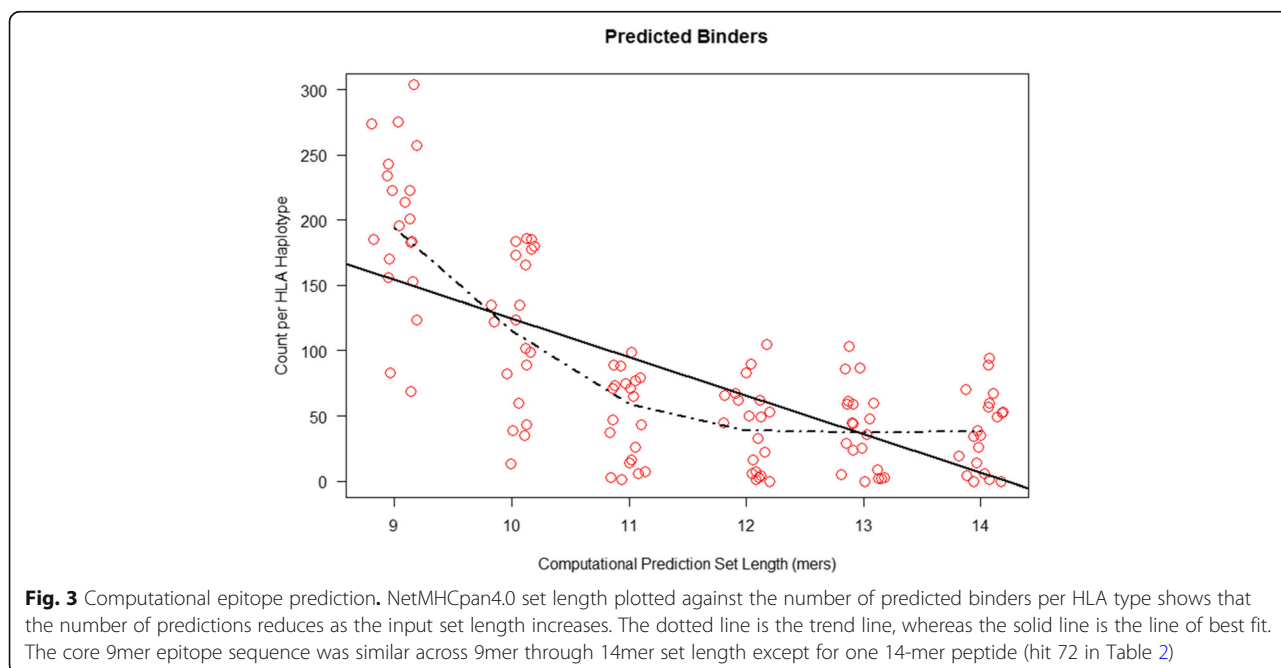
Experimentally mapped peptides that were not predicted by NetMHCpan4.0 as binders. Participant's identifiers beginning with E or L represent early or late time sampling points respectively

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 in the 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 totaling to 3258 assay wells (77%) and leaving only 972 (23%) epitope mapping assays 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 pool, 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%.



**Table 4** Experimental and computational 9mer peptide confusion matrix

	Experimental Positive	Experimental Negative
<b>Computational Positive</b> (≥1 epitope(s) contained in a single experimental peptide sequence)	<b>True Positive (88)</b> (Hits in table 2)	<b>False Positive (37)</b>
<b>Computational Negative</b>	<b>False Negative (5)</b>	<b>True Negative (627)</b>

The total number of peptides experimentally tested were 757 and these are broken down to show the fractions from both the experimental testing and NetMHCpan4.0 computational predictions

### Magnitude of epitope predictions are variable across HLA alleles, HIV-1 proteins and clades

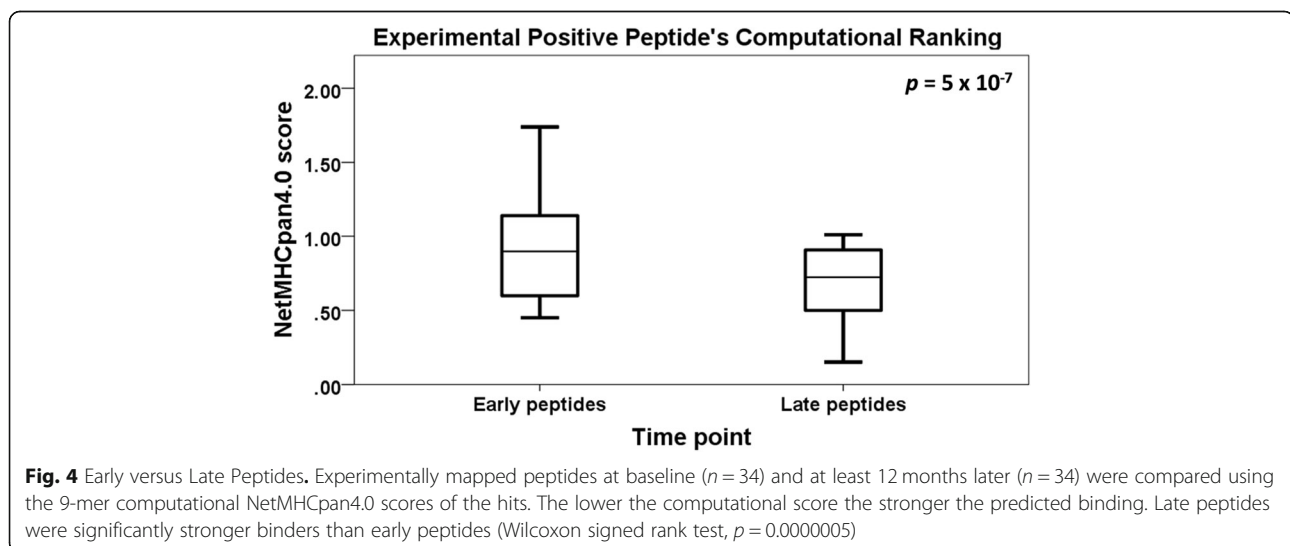
The input HIV-1 subtypes A1 and D consensus whole proteome sequences evaluated for potential 9, 10, 11, 12, 13 and 14-mer binders to the 22 HLA alleles represented in the six patients, varied in the distribution of predicted binders across HIV-1 genes and HLA alleles. All the peptide hits predicted for 10 through 14-mer were also all predicted in the 9-mer set except for two 14-mer peptides. An expected positive correlation for HIV-1 protein length with number of epitopes predicted was observed as illustrated by Spearman's rank order correlation;  $r_s = 0.88$  (Fig. 2, a and b). NetMHCpan4.0 predicted 95% (88/93) (Table 2) of the experimentally mapped peptides as binders and missed 5% (5 out of 93) (Table 3) for the 12-time points of the 6 participants. 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).

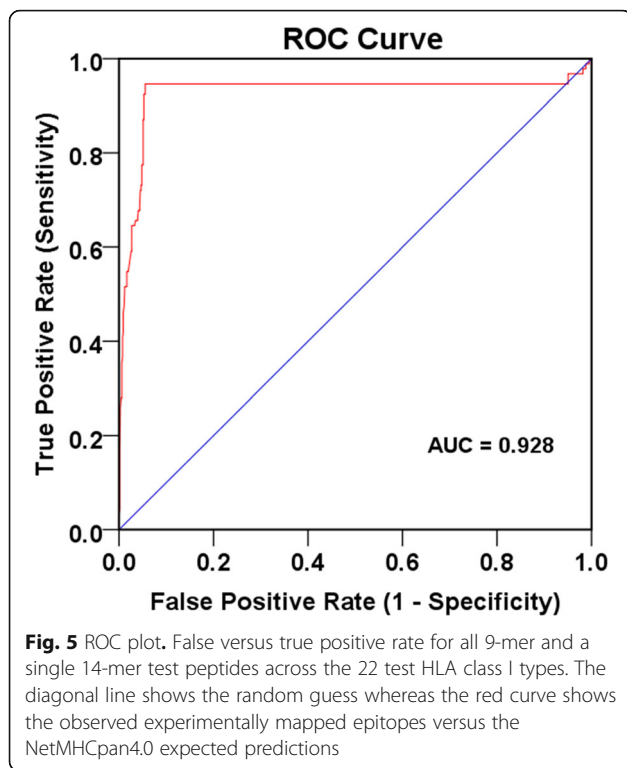
Comparison of the various epitope prediction length set showed that the 9mer setting was ideal for NetMHCpan4.0. The number of predictions were 88, 79, 55, 39, 39 and 37 hits out of 93 for 9, 10, 11, 12, 13 and 14-mer epitopes respectively. Increasing the prediction length from 9mer through 14mer resulted in a smaller number of predicted binders as illustrated in Fig. 3. Since we held the assumption that our wet experimental data was the gold

standard we evaluated the sensitivity and specificity of NetMHCpan4.0. The computational predictor had more predicted binders than those determined by the experimental mapping as presented in the confusion matrix in Table 4. The experimental positive's count also shown in Table 2 under column "Hit No" 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 (Fig. 2, 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 1) 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) (Fig. 4). 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 4) as true positives, false positives, true negatives or false negatives. From the classification the true positive rate (sensitivity) was plotted against the false positive rate (1-specificity) using an ROC curve and the AUC attained reached 0.928 (Fig. 5). 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. 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- $\gamma$  ELISPOT assays we evaluated 757 17mer peptides overlapping by 11 amino acids and covering the whole HIV-1 subtype A1 and D consensus proteomes. Out of the 5 experimentally determined epitopes missed by the algorithm (Table 3), 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- $\gamma$  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 the 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 a single exception, predicted binders between 11 and 14 amino acids included at least one 9mer predicted to bind on its own, suggesting 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- $\gamma$  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.

## 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

## Acknowledgements

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, the late Anthony Kebba who initiated the Rubicon Cohort and Christine Watera who coordinated the recruitment clinic activities.

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.

## Authors' contributions

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

## 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).

## 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.

## 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.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>MRC/UVRI and LSHTM Uganda Research Unit, P. O. Box 49, Plot 51-59 Nakiwogo Road, Entebbe, Uganda. <sup>2</sup>Institute of Human Virology, Abuja, Nigeria. <sup>3</sup>Uganda Virus Research Institute, Entebbe, Uganda.

Received: 28 August 2019 Accepted: 12 February 2020

Published online: 22 February 2020

## References

- Tambunan US, Sipahutar FR, Parikesit AA, Kerami D. Vaccine design for H5N1 based on B- and T-cell epitope predictions. *Bioinform Biol Insights*. 2016;10:27–35.
- Yu K, Petrovsky N, Schonbach C, Koh JY, Brusci V. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med*. 2002;8(3):137–48.
- 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. *J Immunol*. 2008;180(4):2174–86.
- Baden LR, Walsh SR, Seaman MS, Cohen YZ, Johnson JA, Licona JH, Filter RD, Kleinjan JA, Gothering 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;218(4):633–44.
- 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):1890–8.
- 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.
- 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.
- Goodswen SJ, Kennedy PJ, Ellis JT. Discovering a vaccine against neoplasia using computers: is it feasible? *Trends Parasitol*. 2014;30(8):401–11.
- 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):2381–3.
- 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):157–65.
- Currier JR, Robb ML, Michael NL, Marovich MA. Defining epitope coverage requirements for T cell-based HIV vaccines: theoretical considerations and practical applications. *J Transl Med*. 2011;9:212.
- 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. *Nat Immunol*. 2006;7(2):173–8.
- 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. *J Virol*. 2004;78(5):2187–200.
- 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. *J Virol*. 2007;81(5):2440–8.
- 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. *J Virol*. 2006;80(10):4717–28.
- Liu Y, McNeven JP, Holte S, McElrath MJ, Mullins JI. Dynamics of viral evolution and CTL responses in HIV-1 infection. *PLoS One*. 2011;6(1):e15639.
- Rolland M, Heckerman D, Deng W, Rousseau CM, Coovadia H, Bishop K, Goulder PJ, Walker BD, Brander C, Mullins JI. Broad and gag-biased HIV-1 epitope repertoires are associated with lower viral loads. *PLoS One*. 2008;3(1):e1424.
- 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.
- 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. *J Immunol*. 2017;199(9):3360–8.
- 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):83–9.

21. Peters B, Bui HH, Frankild S, Nielson 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 Comput Biol*. 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):358–66.
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):W172–9.
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):2174–81.
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. *Eur J Immunol*. 2005;35(8):2295–303.
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, Wornig 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 Sci*. 2003;12(5):1007–17.
28. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*. 2018;7(1):129–32 e124.
29. DeLisi C, Berzofsky JA. T-cell antigenic sites tend to be amphipathic structures. *Proc Natl Acad Sci U S A*. 1985;82(20):7048–52.
30. He Y, Rappuoli R, De Groot AS, Chen RT. Emerging vaccine informatics. *J Biomed Biotechnol*. 2010;2010:218590.
31. Reche PA, Reinherz EL. Prediction of peptide-MHC binding using profiles. *Methods Mol Biol*. 2007;409:185–200.
32. Roeder 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, Heldebrandt 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):1871–9.
34. Obuku AE, Bugembe DL, Musinguzi K, Watera C, Serwanga J, Ndembu 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 Res Hum Retrovir*. 2016;32(3):237–46.
35. Roederer M, Koup RA. Optimized determination of T cell epitope responses. *J Immunol Methods*. 2003;274(1–2):221–8.
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, Ndembu 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 Res Hum Retrovir*. 2012;28(4):384–92.
38. Serwanga J, Nakiboneka R, Mugaba S, Magambo B, Ndembu 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):1664–72.
39. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, Galvaao-Castro B, Vandamme AM, de Oliveira T. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res*. 2009;37(Web Server issue):W634–42.
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):3797–800.
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 J Radiol*. 2004;5(1):11–8.
43. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics*. 2010;26(6):822–30.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

