frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 December 2011
doi: 10.3389/fimmu.2011.00069

Conservation analysis of dengue virus T-cell epitope-based vaccine candidates using peptide block entropy

Lars Rønn Olsen[1,2], Guang Lan Zhang[1], Derin B. Keskin[3,4], Ellis L. Reinherz[1,3,4] and Vladimir Brusic[1,3]*

[1] Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, MA, USA
[2] Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark
[3] Department of Medicine, Harvard Medical School, Boston, MA, USA
[4] Laboratory of Immunobiology, Dana-Farber Cancer Institute, Boston, MA, USA

Edited by:
Michael Dustin, NYU School of Medicine, USA

Reviewed by:
Christopher E. Rudd, University of Cambridge, UK
Brian M. Baker, University of Notre Dame, USA

***Correspondence:**
Vladimir Brusic, Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Institutes of Medicine 401, 77 Avenue Louis Pasteur, Boston, MA 02118, USA.
e-mail: vladimir_brusic@dfci.harvard.edu

Broad coverage of the pathogen population is particularly important when designing CD8+ T-cell epitope vaccines against viral pathogens. Traditional approaches are based on combinations of highly conserved T-cell epitopes. Peptide block entropy analysis is a novel approach for assembling sets of broadly covering antigens. Since T-cell epitopes are recognized as peptides rather than individual residues, this method is based on calculating the information content of blocks of peptides from a multiple sequence alignment of homologous proteins rather than using the information content of individual residues. The block entropy analysis provides broad coverage of variant antigens. We applied the block entropy analysis method to the proteomes of the four serotypes of dengue virus (DENV) and found 1,551 blocks of 9-mer peptides, which cover 99% of available sequences with five or fewer unique peptides. In contrast, the benchmark study by Khan et al. (2008) resulted in 165 conserved 9-mer peptides. Many of the conserved blocks are located consecutively in the proteins. Connecting these blocks resulted in 78 conserved regions. Of the 1551 blocks of 9-mer peptides 110 comprised predicted HLA binder sets. In total, 457 subunit peptides that encompass the diversity of all sequenced DENV strains of which 333 are T-cell epitope candidates.

Keywords: antigenic diversity, epitope-based vaccines, immunoinformatics, polyvalent vaccines, reverse vaccinology, vaccine informatics

INTRODUCTION

T-cell mediated immunity is a key factor in host responses against human pathogens. It is important for clearance of infection and for anticancer immunity. Peptide-based vaccines offer significant potential advantages in comparison to vaccines using whole proteins or pathogens. These advantages include absence of infectious agents, minimization of negative effects (such as oncogenicity or allergenicity), minimal biological risk (such as reassortment, recombination, or genome integration), ease of production and quality control, and flexibility in inclusion of peptides from multiple molecular targets and their variants (Purcell et al., 2007). Peptides that are recognition targets of CTLs are proposed as key components of the efforts to develop the next generation of vaccines against various diseases including influenza (Brown and Kelso, 2009), HIV (Barouch et al., 2010), and cancers (Pilla et al., 2009). Despite significant research effort, the development of efficient T-cell vaccines has proven difficult (Appay, 2009). The obstacles related to antigenic targeting include diversity of antigenic targets, human leukocyte antigen (HLA) diversity, and availability of peptide targets, i.e., effectiveness of antigen processing and presentation (Brusic and August, 2004; Riemer et al., 2010). Conserved peptides have been studied as targets for epitope-based vaccines in influenza (Tan et al., 2010), variola (Sette et al., 2009), hepatitis C (Yerly et al., 2008), and HIV (Reche et al., 2006; Fischer et al., 2007; Nickle et al., 2007), among others.

Reverse vaccinology (Rappuoli, 2000) provides a conceptual framework where vaccine targets are initially defined from pathogen proteomes using bioinformatics pre-screening, followed by target selection and experimental validation (De Groot and Rappuoli, 2004; Sette and Rappuoli, 2010). The scientific community has long been aware of the need for vaccine strategies which address viral diversity (Hu et al., 1996). In highly variable pathogens, such as HIV, influenza, or dengue, polyvalent vaccines are considered as a solution to viral diversity (Fischer et al., 2007; Morrison et al., 2010).

STRATEGIES FOR DEALING WITH HOST AND VIRAL DIVERSITY IN VACCINE DESIGN

The design of broadly protective T-cell-based vaccines involves identification and selection of vaccine targets composed of conserved antigens containing T-cell epitopes that are both protective and broadly cross-reactive to viral subtypes. The proposed methods of consensus (CON; Gaschen et al., 2002; De Groot et al., 2005), ancestor (ANC; Gao et al., 2005), and center of tree (COT; Nickle et al., 2007) involve assembling individual amino acids into "centralized" consensus sequences of viral proteins that compress

antigenic diversity into a small set of artificial immunogens representative of the virus population. Others, such as the mosaic vaccine design (Fischer et al., 2007), cover viral diversity by assembling naturally occurring peptides identified as T-cell epitopes into artificial proteins, thereby collectively achieving polyvalent coverage. Mosaic vaccines have been tested in preclinical trials with rhesus monkeys showing that these immunogens can induce responses against peptide targets (Barouch et al., 2010; Santra et al., 2010). Common to these methods is a systematic inclusion of highly conserved epitope candidates and exclusion of rare peptides, despite the fact that it has recently been demonstrated that MHC class I epitopes in the flaviviruses have exceptionally low targeting efficiency (i.e., low correlation between MHC binding affinities and conservation of the targeted proteomic regions; Hertz et al., 2011). Furthermore, some low frequency peptides have been shown to be favorable T-cell epitopes in HIV (Rolland et al., 2011). The exclusive focus on highly conserved epitope candidates, for example those maintained in 90% or more of sequences, presents a serious limitation in target selection in dengue virus (DENV) and in other virus species that have multiple serotypes or clades (such as HIV and influenza). In these viruses position many peptides may be intra-clade conserved, but not inter-clade conserved.

A large-scale systematic analysis of peptide conservation and diversity using Shannon entropy calculation and prediction of HLA specificity for conserved peptides in DENV proteins was previously performed for identification of conserved T-cell epitope candidates (Khan et al., 2008). The authors sought to find peptides of nine residues or longer in length, conserved in a minimum of 80% of the 12,404 DENV protein sequences in their dataset. They identified 44 such peptides, of which 34 were conserved in more than 95% of their dataset. They furthermore showed that a subset comprising 34 of the 44 sequences contained 9-mer peptides which were computationally predicted to bind HLA molecules thus representing candidate HLA super type-restricted T-cell epitopes.

Our study extends the analysis of conservation and variability of the DENV proteome. Its focus is the identification of pools of peptides that broadly cover both intra-serotype and inter-serotype diversity. In DENV, the vast majority of experimentally validated epitopes are not conserved across the proteome of all four serotypes. Yet, an efficient vaccine against DENV infection must be protective against all four serotypes. Therefore, the optimal assembly of vaccine targets presents a combinatorial problem where conservation of antigens, diversity of possible HLA interactions, and functional relevance of individual peptides must be assessed. Furthermore, tetravalent formulations (Guy et al., 2010; Murrell et al., 2011) represent main candidate dengue vaccines currently in development. Despite 60 years of research, effective dengue vaccine is not yet available (Murrell et al., 2011). We, therefore, developed and deployed a set of analytical tools to identify, assess, and combine peptide pools suitable for vaccine targeting. It is well-established that the complications associated with secondary infection involve humoral immunity (Halstead and O'rourke, 1977; Halstead et al., 1977), and that immunity against DENV is primarily antibody mediated. Recent research has shown that cellular immunity also plays a role in these complications (Duangchinda et al., 2010), and that cross-reactive DENV-specific T-cells may contribute to the development of dengue hemorrhagic fever (DHF). In this paper we focus on identifying T-cell epitopes, although the block entropy method can be readily used for identification of both linear and discontinuous motifs such as functional B-cell epitopes.

Human $CD8^+$ T-cell epitopes and naturally processed HLA ligands are, with only a few exceptions, 8–11 residues long (Rammensee et al., 1999). $CD4^+$ T-cell epitopes bind HLA molecules mainly through the 9-mer binding core while flanking residues modulate binding (Brown et al., 1993). Therefore, the variability and conservation of antigens should be examined for peptides rather than for individual residues. Block entropy calculations – i.e., calculating information content of a set of aligned sequences – have previously been applied to the analysis of motifs in DNA sequences (Lio et al., 1996). Also, the entropy of 9-mer peptides can be derived from values of individual positions determined using center of the 9-mer calculation (Khan et al., 2008). We applied similar formula for entropy calculations for 8-, 9-, 10-, and 11-mer peptides as blocks. The peptide block entropy method is based on calculating the entropy of blocks of peptides and frequency of individual peptides in a multiple sequence alignment (MSA) of homologous viral protein sequences. The block entropy analysis becomes a useful tool when searching for conserved peptides; for example, a given block may contain only two unique peptides at 50% frequency each, but despite such a modest variability these peptides will not be deemed conserved using the approach that considers conservation of individual amino acids, as used in previous efforts (Fischer et al., 2007; Khan et al., 2008). Because peptide blocks are extracted from a MSA of homologous protein sequences from DENV, a relatively high level of homology of the peptides is expected to be found within an average block. Blocks of homologous peptides are more likely to display similar binding affinity to the HLA class I than randomly assembled sets of non-homologous peptides. However, certain residue variation(s), such as T-cell epitope anchor positions (Falk et al., 1991), can significantly change binding affinity. Similarly, the regions surrounding the block are likely to display inter-sequence homology, suggesting blocks of peptides are more likely to have similar processing characteristics, including proteasomal cleavage and TAP affinity (Martinez et al., 2009), than randomly assembled peptide pools. A block consisting of several epitope candidates with the combined capacity to cover the diversity of all known DENV strains could therefore be a valuable target for prophylactic polyvalent vaccine design.

MATERIALS AND METHODS

VARIABILITY AND CONSERVATION METRICS

The calculation of information content of residues in a MSA of homologous protein sequences is based on the calculation of Shannon entropy (Shannon, 1948):

$$H(x) = -\sum_{i=1}^{I} P_i(x) \log_2(P_i(x)) \quad (1)$$

where H is the entropy, x is the position in the MSA, i represents a given individual amino acid at position x, I is the number of different amino acids on position x, and $P_i(x)$ is the frequency of amino acid i at position x. The conservation of a given position is defined as the frequency of the consensus amino acid (most frequent at a given position).

BLOCK ENTROPY AND CONSERVATION

Shannon entropy can be calculated for each peptide in a block. Each block contains a total of W unique peptides of length l in a dataset of N sequences of length L. We can thus extract $L–l$ blocks, B, of N or fewer unique peptides. The application in conservation analysis is the identification of peptides, which together as a subset, S_w, of W represents a given fraction of W. The formula for calculation of block entropy is:

$$H(B_x) = -\sum_{w=1}^{W} P_w(x)\log_2(P_w(x)) \quad (2)$$

Where $H(B_x)$ is the total entropy of a block of peptides starting at position x, w is a single unique peptide in the space of W unique peptides in block B_x. $P_w(x)$ is the frequency of peptide w at position x.

Four variables are used to classify a block as conserved or not conserved:

1. the minimum number of unique peptides, u, required to reach a pre-defined cumulative frequency in the block in which they are found;
2. the minimum percentage, y_x, of a block that must be covered by the subset of peptides, S_u, for a block to be considered conserved;
3. the maximum allowed fraction, g_x, of peptides containing gaps in the block;
4. the minimum percentage with which each serotype should be covered individually, s_x.

In this analysis we used $u = 5$, $y_x = 0.99$, $g_x = 0.01$, and $s_x = 0.99$. The peptides which are collectively only present in 1% of all sequences $(1 - y_x)$ are unlikely to be stable peptides and many may also be data noise originating from sequencing errors, database entry errors, etc. Assuming that the subset of peptides which collectively occur in less than 1% of known sequences represent variants of low fitness, sequencing errors, or rare variants, we consider that 99% coverage represents a practical threshold for complete conservation. The identification of conserved blocks is combined with the assessment of HLA binding potential for each peptide in each block. Blocks in which all peptides, u, in S_u show similar binding affinity to the same HLA molecule, are classified as "immunofunctionally conserved." Blocks in which not all u in S_u are predicted as HLA binders with the same HLA restriction were discarded.

PREDICTION OF PEPTIDE BINDING TO MHC CLASS I

Human leukocyte antigen binding affinities of peptides in conserved blocks were predicted using NetMHC 3.2 (Lundegaard et al., 2008). Binding affinity to HLA class I was predicted for peptides of nine residues long for the following HLA alleles: HLA-A*0201, HLA-A*03:01, HLA-A*11:01, HLA-A*24:02, HLA-B*07:02, HLA-B*08:01, HLA-B*15:01. These HLA class I alleles were selected for the analysis because NetMHC3.2 predictions of peptide binding to these variants were shown to be highly accurate (Lin et al., 2008). The default thresholds for binding level affinity ($IC_{50} < 500$ nM for weak binders and $IC_{50} < 50$ nM for strong binders) were used for binding classification in this study. Thus a minimum binding affinity of 500 nM was required for a peptide to be considered a potential binder.

DEALING WITH ALIGNMENT GAPS AND AMBIGUOUS CHARACTERS IN THE MSA

Gap insertions in the alignment correspond to insertion or deletion (indel) variation in one or more sequences in the dataset. The DENV diversity is generally caused by substitution mutations rather than indels, but some gaps were observed. Indels of residues can lead to significant change of binding potential or, if both variants are binders, completely different T-cell recognition (Riemer et al., 2010). Therefore, in block entropy based conservation analysis we consider blocks with gaps problematic. In most cases gaps in the alignment were caused by a fraction of the sequences lower than 1% (rare sequences) which were simply removed. If gaps could not be eliminated in this way, the blocks in which more than 1% of the peptides contained gaps were considered too variable and were classified as not conserved. Similarly, peptides containing ambiguous amino acid characters (such as "X") were omitted from the analysis.

SEQUENCE LOGOS

We used sequence logos to visualize the information content (measured in bits) of each position within the blocks (Schneider and Stephens, 1990). Sequence logos are visual representations of the Shannon entropy of the positions within a given sequence. The theoretical maximum entropy of a position in a protein sequence is $\log_2 20 \approx 4.32$ (corresponding to equal representation of all 20 amino acids), so each amino acid on a position can be represented by its fractional information content of the maximum. To generate sequence logos we used WebLogo (Crooks et al., 2004).

BLOCK LOGOS

We designed a logo for visualizing information content of blocks by modifying the sequence logo representation. Sequence logos are very informative about the occurrence of residues on each position, but do not carry valuable information about the frequencies of peptides. Since the theoretical maximum entropy of a block of unlimited size is $\log_2 20^9 \approx 39$ (corresponding to an equal representation of all possible 9-mers), we use the total entropy, $H(B)$, of a block as the maximum bit on the y axis. The information content of each unique peptide, w, in each block, B_x, can be calculated as follows:

$$H(w) = P_w(x)H(B_x) \quad (3)$$

where $H(w)$ is the entropy of peptide w, $P_w(x)$ is the frequency of peptide w, and $H(B_x)$ is the total entropy of the block, B, starting at position, x, in the MSA. The peptides are displayed from most to least frequent starting from the base of the x axis.

DENV SEQUENCES AND T-CELL EPITOPE DATA

The immune epitope database (IEDB; Vita et al., 2010) was queried for known DENV MHC class I binders. For the block entropy analysis we used only complete DENV protein sequences extracted from GenPept (Benson et al., 2010). These sequences were aligned using MAFFT (Katoh and Toh, 2008). Individual protein products were annotated only in a small fraction (roughly 30%) of the polyprotein sequences retrieved from NCBI. The remaining proteomes were annotated using annotation from GenPept reference sequences within the MAFFT alignments. The sequences were deposited in an in-house, publically available database for easy access <cvc.dfci.harvard.edu/flavi> (Olsen et al., 2011). The numbers of sequences classified by protein and serotype are listed in **Table 1**.

Due to sampling bias and natural frequency differences between the four serotypes of DENV, sequences for the serotypes were not found in similar numbers. For example, relatively small number of available DENV4 sequences meant that extra care was required to ensure that DENV4 diversity was covered properly. We therefore adjusted the size of the datasets of DENV2, DENV3, and DENV4 simply by multiplying the datasets of less frequent serotypes to match the most frequent serotype, DENV1. Upon inspection of the adjusted dataset, we concluded that no further significant strain redundancy was present.

Table 1 | Sequence data used in this analysis.

Protein	DENV serotype				Total
	1	2	3	4	
anC	1235	872	739	189	3035
prM	1235	933	742	194	3104
E	1759	1487	1011	409	4666
NS1	1226	912	595	106	2839
NS2A	1241	839	565	105	2750
NS2B	1241	838	565	105	2749
NS3	1214	838	565	105	2722
NS4A	1214	837	566	105	2722
2K	1214	838	566	105	2723
NS4B	1213	838	566	105	2722
NS5	1209	835	566	105	2715
Total	14001	10067	7046	1633	32747

RESULTS AND DISCUSSION

CONSERVATION OF KNOWN T-CELL EPITOPES

Querying the IEDB database (Vita et al., 2010) for experimentally determined DENV $CD8^+$ T-cell epitopes yielded a list of 190 verified 9-mer T-cell epitopes. The average conservation of known T-cell epitopes across the DENV1–4 proteins was 37.13%. Only 18 (10%) of all known epitopes are found in >90% of the DENV1–4 strains (**Figure 1**). Thus only 10% of the known epitopes would be included as potential vaccine targets using a residue-based definition of conservation.

SUMMARY OF CONSERVED PAN-DENV PEPTIDE BLOCKS

We analyzed all blocks of 8, 9, 10, and 11 residues long in the MSA of DENV polyproteins. For each block we calculated the block entropy, the minimum number of peptides needed to cover 99% of a block, the coverage of each of the four DENV serotypes, and also identified the total number of peptides in each block. The conservation of 8-, 9-, 10-, and 11-mer blocks is summarized in **Table 2**.

There are 1,732, 1,551, 1,394, and 1,245 conserved blocks of 8-, 9-, 10-, and 11-mer peptides respectively. Khan et al. (2008) identified 206, 165, 118, and 88 conserved 8-, 9-, 10-, and 11-mer peptides, respectively, by their criteria for conservation (individual peptides conserved in 80% or more sequences). Using peptide block entropy approach to conservation analysis, we found an approximately 10-fold larger conserved target space, which can be examined for potential T-cell epitope candidates. We found the conserved blocks in anC, prM, NS2A, NS2B, NS4A, and 2K proteins, which have previously been considered as too variable for mapping T-cell epitope candidates for cross-protective vaccine constructs.

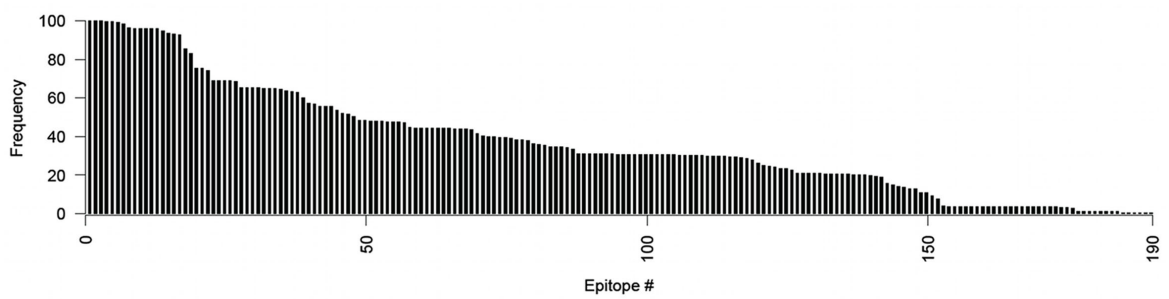

FIGURE 1 | The frequency of the 190 known epitopes sorted from most to least frequent.

By using the conservation thresholds defined in the Section "Materials and Methods," we examined each protein for conservation of blocks. The number of peptides (9-mers) required to cover 99% of the block for each position in the proteome is shown in **Figure 2**. Peptide block conservation relative to protein length was highest in the NS4B and lowest in NS2A proteins. In NS4B, 166 blocks of 9-mers (69.1% of blocks within this protein) were conserved whereas NS2A showed only 8.83% block conservation.

The average entropy for blocks of 9-mer peptides was 1.70 ± 0.71. This is almost double the entropy of individual positions where values larger than 1 indicate highly variable positions (Koo et al., 2009). The entropy of blocks, with only a very few exceptions, where five or less peptides are required to cover 99% of sequences within a block is as high as 2.4 bits (**Figure 3**). This indicates that the block entropy analysis is a robust method, making it suitable for identification of conserved regions of antigenic proteins, where antigenic diversity can be covered by a small number of peptides. This result shows that block entropy analysis is suitable for target selection in polyvalent vaccine formulations.

Table 2 | The total number of blocks which covers 99% of the sequences with five peptides or less, as well as the relative distribution of numbers of peptides in each block.

Peptide length	Total number of blocks	Blocks with 99% coverage with ≤5 peptides	Distribution of number of peptides in blocks				
			1 peptide	2 peptides	3 peptides	4 peptides	5 peptides
8	3393	1732	199	355	388	460	330
9	3392	1551	142	319	341	435	314
10	3391	1394	102	278	295	418	301
11	3390	1245	75	228	266	386	290

This table is a summary of block of 8, 9, 10, and 11-mer peptides.

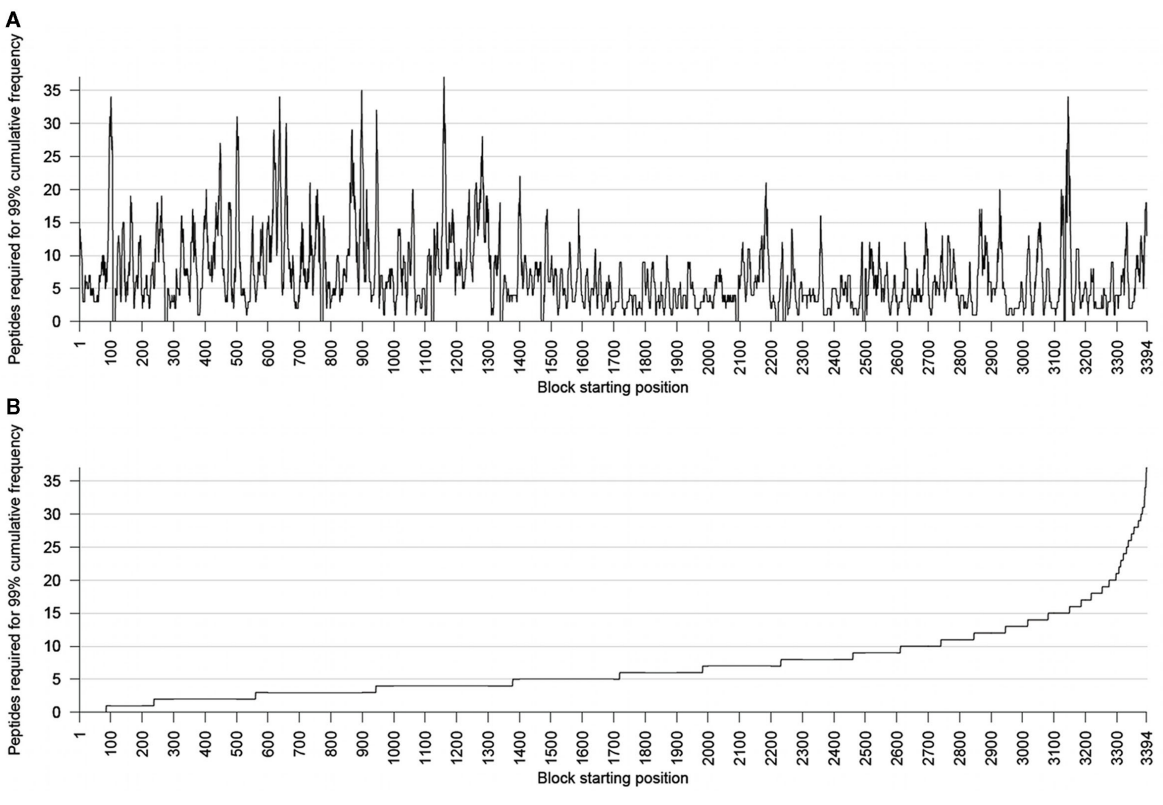

FIGURE 2 | (A) The number of 9-mer peptides required to cover 99% of a block, for each possible start position in the proteome. **(B)** The number of 9-mer peptides required to cover 99% of a block, sorted in increasing order.

BLOCK INFORMATION CONTENT

We examined the information content of individual blocks. **Table 3** shows a representative 9-mer block (position 388 of the MSA of NS3 proteins). We calculated the frequency and information content of each peptide in the block and assessed serotype distribution of these peptides. Five peptides are needed to cover >99% of the sequences within this block, covering approximately 65, 29, 4, 1, and 0.3% of sequenced DENV strains respectively. None of these peptides would have been included in a traditional conservation analysis, in which 80–90% is a typical conservation threshold. This analysis also brings an insight into the effects of threshold selection. If a loose block conservation threshold was used (90%), the three least frequent peptides in the MSA would be excluded and this would exclude DENV4 peptides from the target set. The 99% threshold, on the other hand, would exclude only extremely rare peptides across DENV1–3 serotypes. The peptide number 5 was only found in five strains isolated in Senegal from the late 1960s and three strains from Nigeria from late 1990s. Peptide 5 therefore appears to be a geographically and historically isolated low fitness variant. Peptides 3 and 4 were, however, found in strains isolated almost every year from 1944 to present, and 1983 to present, respectively. Furthermore, peptides 3 and 4 were found distributed across Asia and Australasia and peptide 3 was also observed in Latin America and parts of South America. It is therefore highly likely that strains containing these particular peptides will resurface again given that geographic barriers to spread of DENV are diminishing in the wake of climate changes and increased travel. We, therefore, expect that these strains will continue to spread and

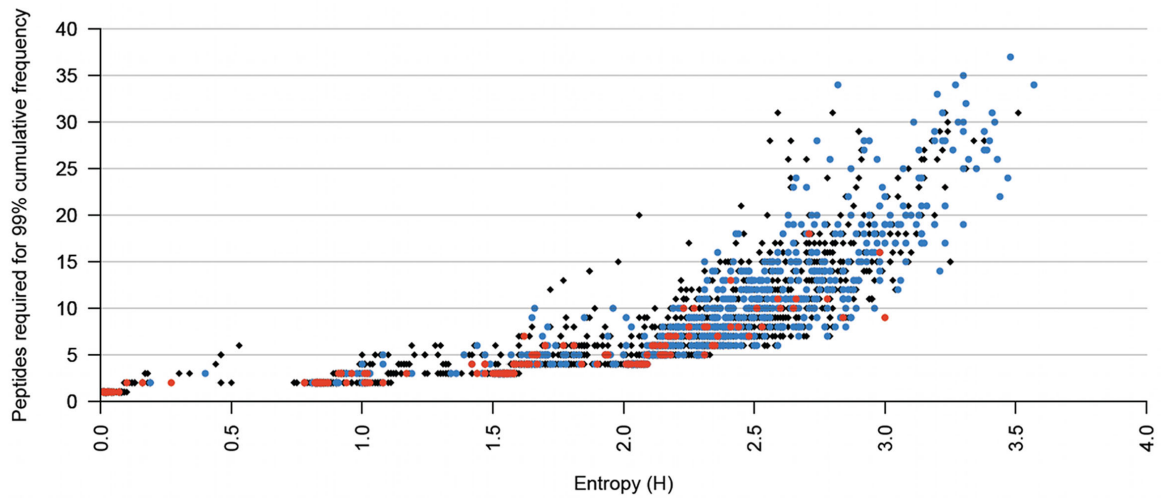

FIGURE 3 | *X, Y* scatter plot of the number of peptides required for 99% coverage of a given block, against the entropy of each given block. The black diamonds correspond to blocks in which no peptides are predicted to bind to the HLA. The blue circles indicate that some, but not all, peptides within that block are predicted to be epitopes. The red squares indicate that all peptides within the block are predicted to bind the same HLA type.

Table 3 | Details of the peptides observed in block 388 of the NS3 protein.

#	Peptide	Frequency	Acc. frequency	Bits	HLA binding (nM)	Serotypes containing peptide (%)
1	KTFDTEYQK	65.03	65.03	0.84	A*03:01(59.69); A*11:01(5.41)	DENV1(99.67) DENV3(99.32)
2	KTFDSEYVK	28.69	93.72	0.37	A*03:01(51.48); A*11:01(7.19)	DENV1(0.08) DENV2(92.64)
3	KTFDTEYPK	3.82	97.54	0.05	A*03:01(26.17); A*11:01(3.35)	DENV4(100)
4	KTFDSEYIK	1.18	98.71	0.02	A*03:01(71.55); A*11:01(8.59)	DENV2(4.19)
5	KTFDTEYTK	0.29	99.01	0.00	A*03:01(37.45); A*11:01(4.65)	DENV2(0.95)
6	KTFDSEYAK	0.29	99.30	0.00	Not predicted	DENV2(0.95)
7	KTFDTEYIK	0.26	99.56	0.00	Not predicted	DENV2(0.90)
8	RTFDTEYQK	0.11	99.67	0.00	Not predicted	DENV1(0.25)
9	KTFETEYQK	0.11	99.78	0.00	Not predicted	DENV3(0.17)
10	KTFDAEYVK	0.07	99.85	0.00	Not predicted	DENV2(0.25)
11	KTFNTEYQK	0.07	99.93	0.00	Not predicted	DENV3(0.34)
12	KTFDTEYQR	0.04	99.96	0.00	Not predicted	DENV3(0.17)
13	KTFDFEYIK	0.04	100	0.00	Not predicted	DENV2(0.12)

Peptide 1 (KTFDTEYQK) is a known ligand for HLA-03, -11, and -31. Furthermore, peptides 2–5 are predicted to have similar binding affinity to peptide 1 for HLA-A3 supertype alleles using MULTIPRED2 (Zhang et al., 2011).

proliferate across the world (Mackenzie et al., 2004; Franco et al., 2010). Modern vaccine development clearly requires variant inclusion beyond target selection resulting from a simple conservation analysis. Block entropy analysis enables identification and further analysis of historical strains, while it is much more difficult to identify relevant low frequency peptides using individual position analysis.

We compared conservation analysis using block entropy with the analysis based on frequency of individual positions. This comparison can be supported by the visualization tools; the sequence logo (Crooks et al., 2004) and our new tool, the Block Logo (**Figure 4**). From the sequence logo one can picture a combinatorial space in which up to eight different peptides maybe present. From the block logo we can see that only two peptides cover 94% while only four peptides within the block show any notable presence.

PREDICTION OF HLA BINDING OF CONSERVED DENV PEPTIDE BLOCKS

Binding affinities were predicted for each of the 5,113 peptides in the 1,551 blocks of conserved 9-mer peptides. If all peptides in a block were predicted to bind to the same HLA type, we consider the block "immuno-functionally conserved" (further defined in Materials and Methods). In total 110 blocks, comprising 333 peptides, were predicted to be immuno-functionally conserved. The distribution of immuno-functionally conserved peptides from different proteins with the number of peptides in each block is shown in **Table 4**.

The antigenic potential differs between individual proteins, as shown by the number of predicted epitope blocks relative to the size of the protein (**Table 5**). A protein that has a high conservation to size ratio is traditionally assumed to have high antigenic potential for vaccine design. Proteins NS3, NS5, NS2B, and anC have high antigenic potential while others, particularly prM and NS2A have low antigenic potential. The 2K protein is predicted to have only one immuno-functionally conserved block, but it is very small in size (23 amino acids), resulting in a high conservation to size ratio.

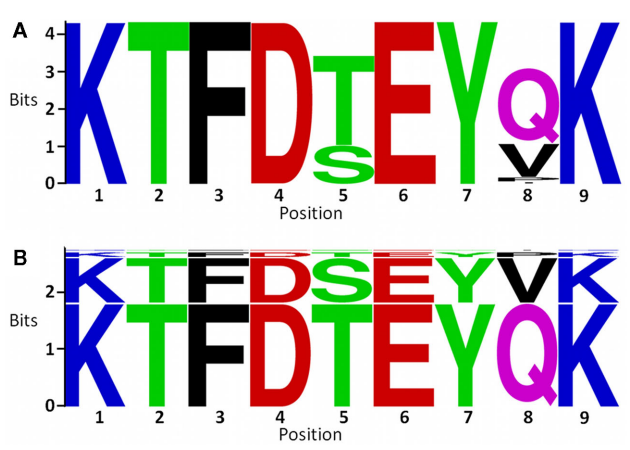

FIGURE 4 | (A) Sequence logo plot of the residues in the block starting at position 388 of the MSA of NS3 protein sequences. The sequence logo was generated using WebLogo (http://weblogo.berkeley.edu/). **(B)** Shows a peptide block logo of the peptides in the same block. The block logo was generated locally using the Block Logo tool. The colors of the amino acids correspond to their chemical properties: polar amino acids (G, S, T, Y, C, Q, and N) are shown in green, basic amino acids (K, R, and H) are shown in blue, acidic amino acids (D and E) are shown in red, and hydrophobic amino acids (A, V, L, I, P, W, F, and M) are shown in black.

As the number of peptides in immune-conserved blocks increases, the higher the entropy. However, blocks that have identical number of peptides can vary in their entropy due to the diversity in the set of peptides making up the remaining 1% of the block excluded from the conserved set. **Figure 3** shows blocks in which some (but not all) peptides are predicted to be epitopes (662 of 3309 peptides), blocks in which all peptides are predicted to bind to the same HLA (151 of 3309 peptides), and blocks predicted to have no HLA binders (2496 of 3309 peptides). The highest concentration of immune-conserved blocks is found in the body of blocks conserved by two, three, and four peptides. We can hypothesize that the immunogenicity of these regions cause higher evolutionary pressure, but that in some regions DENV have only limited space for variability due to possible loss of biological function. This limited variability means that in many cases the peptide variants within a given block may have immunogenic potential with the same HLA restriction and with similar binding affinity, but may require different T-cell clones for immune recognition. Such peptides make excellent targets for polyvalent vaccines for complete coverage.

Although the accuracy of the MHC binding prediction algorithm is high, experimental validation of these epitopes should be performed to ensure first that the peptides are processed and presented to the immune system, and also that they are functional T-cell epitopes. Prediction and experimental validation of 8-, 10-, and 11-mers has yet to be performed. Likewise, prediction and experimental validation of binding can be extended to MHC class II.

COMPRESSING ANTIGENIC DIVERSITY FOR VACCINOLOGY APPLICATIONS

A common approach to designing vaccines against DENV involves polyvalent constructs (Murrell et al., 2011). The block entropy method facilitates the design of polyvalent vaccines by identifying sequences that offer broad coverage of the diversity of all DENV. An example of a broadly neutralizing DENV vaccine design is a tetravalent chimeric live attenuated vaccine developed by Sanofi Pasteur. The vaccine covers all four DENV serotypes and offers protection after three doses are delivered over a 15-months period (Morrison et al., 2010). The long period between the first immunization and protection state presents a limitation, since the risk of DHF from natural secondary infection would be significantly increased between the first and the last dose. Given these limitations of current tetravalent vaccine design (four constructs), and the lack of effective protection with current vaccines, we considered an additional construct and analyzed blocks which require up to five peptides ($w = 5$) to achieve the accumulative coverage of 99%. In regions of consecutive conserved blocks, the peptides can be extended to encompass several conserved blocks. The extended blocks represent the regions in the DENV proteins which can be covered by including five or less peptides in the vaccine construct (**Table 6**). This analysis can be performed for $u > 5$, but such design will include a much

Table 4 | Conserved blocks with all peptides are predicted (NetMHC) to have high HLA binding affinity.

Protein	Block	Class I HLA allele						
		A*02:01	A*03:01	A*11:01	A*24:02	B*07:02	B*08:01	B*15:01
anC	14						3	
	22		5					
	45	4						
	50						3	
E	38			4				
	209	3						3
	211				2			
	237		4	4				
	258	4						
	405						3	
NS1	33	4						
	37					4		
	40	5						
	161			4*				
	191							2
	224					3		
	251							4
	294					2	2	
	313			2				
NS2B	2	5						
	14		5	5				
	29					3		
	46			4				
NS3	43	3						
	48						2	
	95			3				
	130					2		
	183						4	
	204					4		
	217			2				
	218	2						
	234		3					
	238					3		
	258		1					
	259	2						
	263			2				
	275					4		
	280							4
	281	3						
	313	3						
	357	1						
	358			5				
	388		5					
	421					2		
	422			2				
	433	4						
	441	2						

(Continued)

Table 4 | Continued

Protein	Block	Class I HLA allele						
		A*02:01	A*03:01	A*11:01	A*24:02	B*07:02	B*08:01	B*15:01
NS3 (cont'd)	447					4		
	507		5	3				
	528						5	
	531			2				
	536	1						1
	593					4		
NS4A	39				5			
	104							2
	108	3						
2K	5	5						
NS4B	2	5						
	14		5	5				
	29					3		
	46			4				
NS5	52			3				
	81							1
	86		3	3				
	116		4	4				
	117				3			
	181	4						
	209							1
	300		2	2				2
	317			3				
	318	3						3
	325	2						
	340							3
	346					1		
	393			5				
	449			1				
	453						1	
	460		2					
	472	1						
	475	2						
	477						2	2
	489				2			
	490	3						
	570			2				
	587			4				
	679		3*	5				
	757	5						
	759							4
	760	4			4			4
	765						1	
	767						1	
	772						3	
	849			3				
	850	3						

The number in the cells corresponds to the number of peptides needed to cover 99% of the block in which it is found. For the numbers marked with an asterisks (*), the least predominant peptide(s) were not predicted to be an MHC binder, but the remaining peptides cover at least 95%.

Table 5 | The ratio of conservation to size of each DENV protein is shown.

Protein	Fraction of proteome (%)	Immuno-functionally conserved blocks (%)	Immuno-functional conservation: size ratio
anC	3.36	3.64	1.08
prM	4.89	0.00	0.00
E	14.59	7.27	0.50
NS1	10.38	9.09	0.88
NS2A	6.43	0.00	0.00
NS2B	3.83	4.55	1.19
NS3	18.25	29.09	1.59
NS4A	3.74	2.73	0.73
2K	0.68	0.91	1.34
NS4B	7.34	4.55	0.62
NS5	26.50	38.18	1.44

The conservation to size ratio was calculated by dividing the number of blocks predicted to be immuno-functionally conserved, divided by the relative size (in %) of each protein.

larger number of constructs in the polyvalent vaccine. While the main focus of this work is on the selection of T-cell epitope targets for DENV vaccine development, it also provides a method which can be used as input for experimental design of other immunological studies, such as examining immunodominance, competitive epitope binding, and detrimental cross-reactivity.

EXPERIMENTAL SUPPORT OF PRINCIPLE

We examined IEDB and current literature for examples of experimentally validated epitopes and compared them with our results. We found four examples where predictions correspond to experimental data (**Table 7A**) and two examples where predictions did not match fully the experimental data (**Table 7B**).

The blocks presented in **Table 7A** consist of four to six peptides with accumulated minimum frequency of 98% that were both predicted and experimentally validated HLA binders. The blocks presented in **Table 7B** consist of three and seven peptides respectively. Common for these two blocks is that one or more peptides experimentally shown to be HLA binders were not predicted to be binders. These two blocks, although potentially useful in a polyvalent vaccine construct, where not identified as universally binding by the prediction algorithm. The implication is that the blocks where majority or all of peptides are identified as potential binders should be experimentally validated. Conversely, the blocks where majority of peptides are identified as non-binders are less likely to be experimental T-cell epitopes.

CONCLUSION

The analysis of conservation of DENV antigens should consider both the pan-DENV antigenic diversity and the diversity between DENV serotypes. Furthermore, functional properties, such as HLA binding potential are essential for the assessment of immunogenic potential of antigens. DENV conservation analysis in previous studies was based on the traditional approach, in which a peptide is classified as conserved or not conserved based on analysis of each individual amino acid along with an arbitrary frequency threshold (typically 90% or higher). The premise of traditional conservation analysis of vaccine targets is that conserved epitope candidates are more likely to confer cross-protection between pathogen variants. We argue that variant inclusion is important for polyvalent coverage, since the array of factors making a peptide immunogenic is far too complex to assume that conserved predicted binders are automatically the best immunogens. Our systematic approach, that deploys analysis of conservation of blocks of peptides, has produced a 10-fold larger number of potential DENV T-cell epitope targets than the traditional approach. Similar to a previous benchmark study (Khan et al., 2008), our method also enables vaccine target discovery that considers both the conservation of antigens and the immunogenic potential of these peptides. The peptide blocks determined in our study can be used to inform and focus the design of experimental studies of polyvalent dengue vaccines. Furthermore, our approach is applicable to any variable virus such as HIV, influenza, or Hepatitis C. It is also applicable to broader vaccine approaches such as identification of shared peptide targets across major *Flavivirus* pathogens.

The block entropy analysis is an informed strategy for achieving broad strain coverage in vaccine design with the inclusion of significant but less frequent variants. Central to this approach is the fact that T-cell epitopes are recognized as peptides rather than single residues, and should therefore be analyzed as such. For example, the analysis of 9-mer blocks enables characterization of a set of peptides which collectively can be considered as conserved for immunological applications. In this study, we based the assessment of immunological potential using predictions of peptide binding to seven common HLA class I molecules for which prediction algorithms have been validated (Lin et al., 2008). Some of the conserved blocks contain peptides which are all predicted to bind the same HLA allele – these blocks are considered to be immuno-functionally conserved. The peptide block thus becomes a unit of analysis for building combinatorial vaccine formulations with broad coverage of both pathogen variants and diverse HLA haplotypes. This analysis provides a reasonably sized set of targets that can be experimentally validated, for example by mass spectrometry (Reinhold et al., 2010).

The premise for the concept of immuno-functional conservation is not only that all peptides in a block bind to MHC with the same affinity and HLA restriction, but also that there is enough redundancy in CTLs so that each epitope/MHC complex may elicit an equally strong T-cell response. However, immunodominance of certain epitopes and some T-cell receptors (TCRs; Nikolich-Zugich et al., 2004) can lead to an uneven response to antigens upon vaccination and thus incomplete strain coverage upon challenge. High intra-block homology could allow for a population of CTLs to recognize all epitope in a block, which may also favor the concept of including entire immuno-functionally conserved blocks in a polyvalent vaccine construct. Hence, predicting the cross-protective capacity of a peptide-based vaccine gets more difficult as the number of T-cell epitope

Table 6 | Extended block sequences for string 1.

Extended block ID	Protein	Position	Sequence
1,1	anC	9–69	PPFNMLKRERNRVSTGSQLAKRFSKGLFSGQGPMKLVMAFIAFLRFLAIPPTAGILKRWG
1,2	anC	81–91	GFRKEIGRML
1,3	anC	113–127	FHLTTRNGEPHMIV
1,4	prM	15–27	QERGKSLLFKTA
1,5	prM	30–47	NMCTLIAMDLGELCEDT
1,6	prM	58–79	EPEDIDCWCNLTSTWVTYGTC
1,7	prM	82–120	GEHRRDKRSVALVPHVGMGLETRTETWMSSEGAWKHAQ
1,8	prM	121–132	VETWALRHPGF
1,9	prM	165–213	MRCVGVGNRDFVEGLSGATWVDVVLEHGSCVTTMAKNKPTLDFELIKT
1,10	E	67–79	TTDSRCPTQGEA
1,11	E	94–118	FVDRGWGNGCGLFGKGSLVTCAKF
1,12	E	181–198	LECSPRTGLDFNEMVLL
1,13	E	202–219	KSWLVHKQWFLDLPLPW
1,14	E	233–271	DLLVTFKTAHAKKQEVVVLGSQEGAMHTALTGATEIQT
1,15	E	276–292	IFAGHLKCRLKMDKLT
1,16	E	309–320	EVAETQHGTVL
1,17	E	363–377	VNIEAEPPFGDSYI
1,18	E	392–427	KGSSIGKMFEATARGARRMAILGDTAWDFGSVGGV
1,19	E	438–452	FGTAYGVLFSGVSW
1,20	E	460–471	LLTWLGLNSRS
1,21	NS1	8–48	GKELKCGSGIFVTDEVHTWTEQYKFQADSPSKLASAIQKA
1,22	NS1	51–70	GVCGIRSVTRLENIMWKQI
1,23	NS1	111–121	YSWKTWGKAK
1,24	NS1	151–170	WEVEDYGFGVFTTNIWLKL
1,25	NS1	177–210	CDHRLMSAAIKDSKAVHADMGYWIESSKNQTWQ
1,26	NS1	211–222	EKASLIEVKTC
1,27	NS1	223–244	WPKSHTLWSNGVLESEMIIPK
1,28	NS1	250–259	SQHNYRPGY
1,29	NS1	260–276	TQTAGPWHLGKLELDF
1,30	NS1	289–337	CGNRGPSLRTTTASGKLIHEWCCRSCTMPPLRFRGEDGCWYGMEIRPL
1,31	NS2A	90–104	FLRKLTSRETALMV
1,32	NS2A	217–272	SWPLNEGIMAVGMVSILASALLKNDIPMTGPLVAGGLLTVCYVLSGSSADLSLEK
1,33	NS2B	62–72	QAEISGSSPI
1,34	NS2B	76–91	QQEDGSMSIKNEEEE
1,35	NS2B	92–107	MLTILIRTGLLVISG
1,36	NS2B	129–141	AGVLWDVPSPPP
1,37	NS3	29–40	FGYSQIGAGVY
1,38	NS3	41–59	EGVFHTMWHVTRGSVICH
1,39	NS3	60–84	GGRLEPSWASVKKDLISYGGGWRL
1,40	NS3	87–110	WKEGEEVQVIAVEPGKNPRAVQT
1,41	NS3	118–140	TGTIGAIALDFKPGTSGSPIIN
1,42	NS3	141–168	KGKVVGLYGNGVVTKSGDYVSAITQAE
1,43	NS3	182–321	FRKRRLTIMDLHPGAGKTKRYLPAIVREALKRRLRTLILAPTRVVAAEMEEALRGLPIRYQTPAVKSEHTGR EIVDLMCHATFTMRLLSPVRVPNYNLIIMDEAHFTDPASIAARGYISTRVGMGEAAAIFMTATPPGS
1,44	NS3	322–348	DPFPQSNSPIQDEERDIPERSWNTGF
1,45	NS3	353–397	FKGKTVWFVPSIKAGNDIANCLRKNGKKVIQLSRKTFDTEYQKT
1,46	NS3	398–465	LNDWDFVVTTDISEMGANFRADRVIDPRRCLKPVILTDGPERVILAGPMPVTVASAAQRRGRIGRNP
1,47	NS3	466–557	NENDQYIYMGQPLNNDEDHAHWTEAKMLLDNINTPEGIIPALFEPEREKSAAIDGEYRLRGEARKTF VELMRRGDLPVWLSYKVASAGISY
1,48	NS3	567–618	RNNQILEENMDVEIWTKEGERKKLRPRWLDARVYSDPLALKEFKEFAAGRK
1,49	NS4A	3–13	NLITEMGRLP
1,50	NS4A	20–34	KNALDNIVMLHTTE

(Continued)

Table 6 | Continued

Extended block ID	Protein	Position	Sequence
1,51	NS4A	38–66	AYQHALSELPETLETLLLLALLGAMTAG
1,52	NS4A	98–141	IQPHWIAASIILEFFLMVLLIPEPEKQRTPQDNQLTYVVIAIL
1,53	2K	22–36	NEMGFLETTKKDLG
1,54	NS4B	27–111	LDVDLRPASAWTLYAVATTVLTPMLRHSIENSSVNVSLTAIANQAAVLMGLDKGWPLSKMDLGVPLLAL GCYSQVNPLTLTAAV
1,55	NS4B	115–242	AHYAIIGPGLQAKATREAQKRTAAGIMKNPTVDGITVIDLEPIPYDPKFEKQLGQVMLLVLCVTQVLLMR TTWALCEALTLATGPITTLWEGNPGRFWNTTIAVSMANIFRGSYLAGAGLAFSLIKN
1,56	NS5	3–17	QGETLGEKWKRQLN
1,57	NS5	26–48	YKKSGIQEVDRTEAKEGLKRGE
1,58	NS5	50–70	HHAVSRGSAKLRWFVERNLV
1,59	NS5	71–133	PKGKVVDLGCGRGGWSYYCAGLKKVTEVKGYTKGGPGHEEPIPMATYGWNLVKLHSGVDVFY
1,60	NS5	134–171	PPEKCDTLLCDIGESSSNPTIEEGRTLRVLKMVEPWL
1,61	NS5	174–190	QFCIKVLNPYMPTVIE
1,62	NS5	199–243	GGMLVRNPLSRNSTHEMYWVSNASGNIVSSVNMISRMLINRFTM
1,63	NS5	247–266	ATYEKDVDLGAGTRSVSTE
1,64	NS5	269–282	PDMDIIGKRIEKI
1,65	NS5	288–363	SWHYDQENPYKTWAYHGSYETKQTGSASSMVNGVVKLLTKPWDVIPMVTQMAMTDTTPFGQQR VFKEKVDTRTPR
1,66	NS5	388–397	PRLCTREEF
1,67	NS5	399–424	KVRSNAAIGAVFTDENKWKSAREAV
1,68	NS5	435–518	ERALHQEGKCETCVYNMMGKREKKLGEFGKAKGSRAIWYMWLGARFLEFEALGFLNEDHWFSREN SLSGVEGEGLHKLGYILR
1,69	NS5	524–548	GGAMYADDTAGWDTRITEDDLQNE
1,70	NS5	564–583	AIFKLTYQNKVVKVLRPTP
1,71	NS5	584–626	GTVMDIISRKDQRGSGQVGTYGLNTFTNMEAQLIRQMEGEGV
1,72	NS5	650–672	ERLKRMAISGDDCVVKPIDDRF
1,73	NS5	674–720	ALTALNDMGKVRKDIPQWEPSKGWKDWQQVPFCSHHFHELIMKDGR
1,74	NS5	721–783	LVVPCRNQDELIGRARISQGAGWSLRETACLGKAYAQMWSLMYFHRRDLRLAANAICSAVPV
1,75	NS5	784–816	WVPTSRTTWSIHAHHQWMTTEDMLTVWNRVWI
1,76	NS5	817–826	ENPWMEDKT
1,77	NS5	832–861	DVPYLGKREDQWCGSLIGLTSRATWAKNI
1,78	NS5	878–888	DYMPSMKRFR

The extended block ID is composed of the peptide number (ranging from the most frequent in the given block to the least frequent), followed the block number in order of starting position from the N-terminal end of the proteome to the C-terminal end. For the extended blocks in the remaining four strings and detailed mapping of predicted T-cell epitopes, see Table S1 in Supplementary Material. The definition of the strings and the corresponding extended blocks is described in Section "Compressing Antigenic Diversity for Vaccinology Applications."

candidates increases. Furthermore, the larger the combinatorial space needed to cover the diversity of the four serotypes; the more complex the task to combine all of the epitopes in one vaccine without compromising its efficacy. Considering these factors, we choose to include all blocks in which five peptides or less cover 99% of the block. This number maybe subject to adjustment after proper experimental validation of the epitope pools.

Applying block entropy analysis to the proteomes of DENV1–4, yielded 1,732, 1,551, 1,394, and 1,245 conserved blocks of 8-, 9-, 10-, and 11-mer peptides respectively, as opposed to the results of the benchmark study (Khan et al., 2008) which yielded 206, 165, 118, and 88 conserved 8-, 9-, 10-, and 11-mer peptides respectively, using the traditional criteria for conservation. Of the 1,551 blocks of 9-mer peptides, 110 blocks, consisting of 333 peptides, were predicted to be immuno-functionally conserved, based on their predicted binding affinity to HLA class I, which can form the basis of a T-cell-based polyvalent vaccine against DENV. The method presented here can be readily applied to other relevant viral pathogens such as influenza, HIV, or HPV, as well as extended to encompass MHC class II epitope candidates and functional B-cell epitopes.

ACKNOWLEDGMENTS

This work was supported by NIH grant U01 AI 90043 (Guang Lan Zhang, Derin B. Keskin, Ellis L. Reinherz, and Vladimir Brusic). Lars Rønn Olsen was supported by a number of Danish student grants (Otto Mønsteds Foundation; Rudolph Als Foundation; Civil Engineer Frants Alling's Scholarship; Julie

Table 7 | (A) Examples of experimental evidence supporting the application of the block entropy approach to achieve broad coverage by including homologous peptide blocks in polyvalent vaccine constructs. In the four examples below, a high percentage of pan-DENV population coverage was achieved by including only blocks of peptides that have all been predicted, as well as experimentally validated, to bind HLA. Predictions for B*55:02 were done using netMHCpan 2.4 (PMID: 17726526). **(B)** Two examples of blocks of experimentally validated epitopes, for which the block entropy approach failed to account due to epitope predictions inconsistent with the experimental findings of the respective authors.

Peptides	Reference	Experimental HLA	Predicted affinity (nM)	Predicted HLA	Block position	Pan-DENV peptide frequency	Protein
A							
MLLALIAVL	16493038	A*02:01	10	A*02:01	2151	24.86	NS4A
LLLTLLATV	16493038	A*02:01	9	A*02:01	2151	24.79	NS4A
LLLGLMILL	16493038	A*02:01	14	A*02:01	2151	23.29	NS4A
MLVALLGAM	16493038	A*02:01	300	A*02:01	2151	25.23	NS4A
					Accumulated frequency:	98.17	
TPEGIIPSM	7529799	B*35:01	49	B*35:01	1977	25.78	NS3
TPEGIIPTL	7529799	B*35:01	345	B*35:01	1977	25.73	NS3
TPEGIIPAL	7544398	B*35	108	B*35:01	1977	48.19	NS3
TPEGIIPSL	7544398	B*35	836	B*35:01	1977	0.15	NS
					Accumulated frequency:	99.85	
KYTDRKWCF	16517753	A*24	21	A*24:02	2033	23.18	NS3
NYADRKWCF	16517753	A*24	32	A*24:02	2033	7.67	NS3
QYSDRRWCF	16517753	A*24	17	A*24:02	2033	24.89	NS3
SYKDREWCF	16517753	A*24	18	A*24:02	2033	24.99	NS3
NYADRRWCF	11709777	A*24	23	A*24:02	2033	17.62	NS3
					Accumulated frequency:	98.35	
KPWDIIPMV	17626101	B*55:02	149	B*55:02	2828	0.04	NS5
KPWDVIPMV	17626101	B*55:02	192	B*55:02	2828	51.09	NS5
KPWDVLPMV	17626101	B*55:02	158	B*55:02	2828	0.30	NS5
KPWDVLPTV	17626101	B*55:02	107	B*55:02	2828	7.53	NS5
KPWDVVPMV	17626101	B*55:02	139	B*55:02	2828	38.73	NS5
KPWDVVPTV	17626101	B*55:02	112	B*55:02	2828	1.65	NS5
					Accumulated frequency:	99.34	
B							
FLDLPLPWT	16493038	A*02	64	A*02:01	493	50.13	E
FLDLPLPWL	16493038	A*02	11	A*02:01	493	22.7	E
FFDLPLPWT	16493038	A*02:01	13995	A*02:01	493	23.83	E
					Accumulated frequency:	96.66	
GTSGSPIIDKK	12808447	A*11	17	A*11:01	1611	9.36	NS3
GTSGSPIVDRK	12808447	A*11	20	A*11:01	1611	14.2	NS3
GTSGSPIVDKK	12808447	A*11	18	A*11:01	1611	2.26	NS3
GTSGSPIADKK	12808447	A*11	22	A*11:01	1611	0.04	NS3
GTSGSPIVNRE	12808447	A*11	18023	A*11:01	1611	24.97	NS3
GTSGSPIINRE	12808447	A*11	18156	A*11:01	1611	23.26	NS3
GTSGSPIINRK	12808447	A*11	16	A*11:01	1611	21.53	NS3
					Accumulated frequency:	95.62	

Damm's; Rebild National Park Society, Inc.; Inge and Jørgen Larsen's Memorial Scholarship; Mayor Niels Albrechtsen's Scholarship; Danish Society of Engineers Scholarship; and Oticon Foundation).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/t_cell_biology/10.3389/fimmu.2011.00069/abstract

REFERENCES

Appay, V. (2009). 25 years of HIV research!. . .and what about a vaccine? *Eur. J. Immunol.* 39, 1999–2003.

Barouch, D. H., O'brien, K. L., Simmons, N. L., King, S. L., Abbink, P., Maxfield, L. F., Sun, Y. H., La Porte, A., Riggs, A. M., Lynch, D. M., Clark, S. L., Backus, K., Perry, J. R., Seaman, M. S., Carville, A., Mansfield, K. G., Szinger, J. J., Fischer, W., Muldoon, M., and Korber, B. (2010). Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nat. Med.* 16, 319–323.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers,

E. W. (2010). GenBank. *Nucleic Acids Res.* 38, D46–D51.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., and Wiley, D. C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364, 33–39.

Brown, L. E., and Kelso, A. (2009). Prospects for an influenza vaccine that induces cross-protective cytotoxic T lymphocytes. *Immunol. Cell Biol.* 87, 300–308.

Brusic, V., and August, J. T. (2004). The changing field of vaccine development in the genomics era. *Pharmacogenomics* 5, 597–600.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.

De Groot, A. S., Marcon, L., Bishop, E. A., Rivera, D., Kutzler, M., Weiner, D. B., and Martin, W. (2005). HIV vaccine development by computer assisted design: the GAIA vaccine. *Vaccine* 23, 2136–2148.

De Groot, A. S., and Rappuoli, R. (2004). Genome-derived vaccines. *Expert Rev. Vaccines* 3, 59–76.

Duangchinda, T., Dejnirattisai, W., Vasanawathana, S., Limpitikul, W., Tangthawornchaikul, N., Malasit, P., Mongkolsapaya, J., and Screaton, G. (2010). Immunodominant T-cell responses to dengue virus NS3 are associated with DHF. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16922–16927.

Falk, K., Rotzschke, O., Stevanovic, S., Jung, G., and Rammensee, H. G. (1991). Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351, 290–296.

Fischer, W., Perkins, S., Theiler, J., Bhattacharya, T., Yusim, K., Funkhouser, R., Kuiken, C., Haynes, B., Letvin, N. L., Walker, B. D., Hahn, B. H., and Korber, B. T. (2007). Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* 13, 100–106.

Franco, C., Hynes, N. A., Bouri, N., and Henderson, D. A. (2010). The dengue threat to the United States. *Biosecur. Bioterror.* 8, 273–276.

Gao, F., Weaver, E. A., Lu, Z., Li, Y., Liao, H. X., Ma, B., Alam, S. M., Scearce, R. M., Sutherland, L. L., Yu, J. S., Decker, J. M., Shaw, G. M., Montefiori, D. C., Korber, B. T., Hahn, B. H., and Haynes, B. F. (2005). Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. *J. Virol.* 79, 1154–1163.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., and Korber, B. (2002). Diversity considerations in HIV-1 vaccine selection. *Science* 296, 2354–2360.

Guy, B., Saville, M., and Lang, J. (2010). Development of Sanofi Pasteur tetravalent dengue vaccine. *Hum. Vaccin.* 6, 696–705.

Halstead, S. B., and O'rourke, E. J. (1977). Dengue viruses and mononuclear phagocytes. I. Infection enhancement by non-neutralizing antibody. *J. Exp. Med.* 146, 201–217.

Halstead, S. B., O'rourke, E. J., and Allison, A. C. (1977). Dengue viruses and mononuclear phagocytes. II. Identity of blood and tissue leukocytes supporting in vitro infection. *J. Exp. Med.* 146, 218–229.

Hertz, T., Nolan, D., James, I., John, M., Gaudieri, S., Phillips, E., Huang, J. C., Riadi, G., Mallal, S., and Jojic, N. (2011). Mapping the landscape of host-pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. *J. Virol.* 85, 1310–1321.

Hu, D. J., Dondero, T. J., Rayfield, M. A., George, J. R., Schochetman, G., Jaffe, H. W., Luo, C. C., Kalish, M. L., Weniger, B. G., Pau, C. P., Schable, C. A., and Curran, J. W. (1996). The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention. *JAMA* 275, 210–216.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* 9, 286–298.

Khan, A. M., Miotto, O., Nascimento, E. J., Srinivasan, K. N., Heiny, A. T., Zhang, G. L., Marques, E. T., Tan, T. W., Brusic, V., Salmon, J., and August, J. T. (2008). Conservation and variability of dengue virus proteins: implications for vaccine design. *PLoS Negl. Trop. Dis.* 2, e272. doi:10.1371/journal.pntd.0000272

Koo, Q. Y., Khan, A. M., Jung, K. O., Ramdas, S., Miotto, O., Tan, T. W., Brusic, V., Salmon, J., and August, J. T. (2009). Conservation and variability of West Nile virus proteins. *PLoS ONE* 4, e5352. doi:10.1371/journal.pone.0005352

Lin, H. H., Ray, S., Tongchusak, S., Reinherz, E. L., and Brusic, V. (2008). Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol.* 9, 8. doi:10.1186/1471-2172-9-8

Lio, P., Politi, A., Buiatti, M., and Ruffo, S. (1996). High statistics block entropy measures of DNA sequences. *J. Theor. Biol.* 180, 151–160.

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. (2008). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* 36, W509–W512.

Mackenzie, J. S., Gubler, D. J., and Petersen, L. R. (2004). Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. *Nat. Med.* 10, S98–S109.

Martinez, A. N., Tenzer, S., and Schild, H. (2009). T-cell epitope processing (the epitope flanking regions matter). *Methods Mol. Biol.* 524, 407–415.

Morrison, D., Legg, T. J., Billings, C. W., Forrat, R., Yoksan, S., and Lang, J. (2010). A novel tetravalent dengue vaccine is well tolerated and immunogenic against all 4 serotypes in *Flavivirus*-naive adults. *J. Infect. Dis.* 201, 370–377.

Murrell, S., Wu, S. C., and Butler, M. (2011). Review of dengue virus and the development of a vaccine. *Biotechnol. Adv.* 29, 239–247.

Nickle, D. C., Rolland, M., Jensen, M. A., Pond, S. L., Deng, W., Seligman, M., Heckerman, D., Mullins, J. I., and Jojic, N. (2007). Coping with viral diversity in HIV vaccine design. *PLoS Comput. Biol.* 3, e75. doi:10.1371/journal.pcbi.0030075

Nikolich-Zugich, J., Slifka, M. K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* 4, 123–132.

Olsen, L. R., Zhang, G. L., Reinherz, E. L., and Brusic, V. (2011). FLAVIdB: a data mining system for knowledge discovery in flaviviruses with direct applications in immunology and vaccinology. *Immunome Res.* 8, 1.

Pilla, L., Rivoltini, L., Patuzzo, R., Marrari, A., Valdagni, R., and Parmiani, G. (2009). Multipeptide vaccination in cancer patients. *Expert Opin. Biol. Ther.* 9, 1043–1055.

Purcell, A. W., Mccluskey, J., and Rossjohn, J. (2007). More than one reason to rethink the use of peptides in vaccine design. *Nat. Rev. Drug Discov.* 6, 404–414.

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanovic, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50, 213–219.

Rappuoli, R. (2000). Reverse vaccinology. *Curr. Opin. Microbiol.* 3, 445–450.

Reche, P. A., Keskin, D. B., Hussey, R. E., Ancuta, P., Gabuzda, D., and Reinherz, E. L. (2006). Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. *Med. Immunol.* 5, 1.

Reinhold, B., Keskin, D. B., and Reinherz, E. L. (2010). Molecular detection of targeted major histocompatibility complex I-bound peptides using a probabilistic measure and nanospray MS(3) on a hybrid quadrupole-linear ion trap. *Anal. Chem.* 82, 9090–9099.

Riemer, A. B., Keskin, D. B., Zhang, G., Handley, M., Anderson, K. S., Brusic, V., Reinhold, B., and Reinherz, E. L. (2010). A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers. *J. Biol. Chem.* 285, 29608–29622.

Rolland, M., Frahm, N., Nickle, D. C., Jojic, N., Deng, W., Allen, T. M., Brander, C., Heckerman, D. E., and Mullins, J. I. (2011). Increased breadth and depth of cytotoxic T lymphocytes responses against HIV-1-B Nef by inclusion of epitope variant sequences. *PLoS ONE* 6, e17969. doi:10.1371/journal.pone.0017969

Santra, S., Liao, H. X., Zhang, R., Muldoon, M., Watson, S., Fischer, W., Theiler, J., Szinger, J., Balachandran, H., Buzby, A., Quinn, D., Parks, R. J., Tsao, C. Y., Carville, A., Mansfield, K. G., Pavlakis, G. N., Felber, B. K., Haynes, B. F., Korber, B. T., and Letvin, N. L. (2010). Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. *Nat. Med.* 16, 324–328.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100.

Sette, A., Grey, H., Oseroff, C., Peters, B., Moutaftsi, M., Crotty, S., Assarsson, E., Greenbaum, J., Kim, Y., Kolla, R., Tscharke, D., Koelle, D., Johnson, R. P., Blum, J., Head, S., and Sidney, J. (2009). Definition of epitopes and antigens recognized by vaccinia specific immune responses: their conservation in variola virus sequences, and use as a model system to study complex pathogens. *Vaccine* 27(Suppl. 6), G21–G26.

Sette, A., and Rappuoli, R. (2010). Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* 33, 530–541.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423; 623–656.

Tan, P. T., Heiny, A. T., Miotto, O., Salmon, J., Marques, E. T., Lemonnier, F., and August, J. T. (2010). Conservation and diversity of influenza A H1N1 HLA-restricted T cell epitope candidates for epitope-based vaccines. *PLoS ONE* 5, e8754. doi: 10.1371/journal.pone.0008754

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2010). The immune epitope database 2.0. *Nucleic Acids Res.* 38, D854–D862.

Yerly, D., Heckerman, D., Allen, T., Suscovich, T. J., Jojic, N., Kadie, C., Pichler, W. J., Cerny, A., and Brander, C. (2008). Design, expression, and processing of epitomized hepatitis C virus-encoded CTL epitopes. *J. Immunol.* 181, 6361–6370.

Zhang, G. L., Deluca, D. S., Keskin, D. B., Chitkushev, L., Zlateva, T., Lund, O., Reinherz, E. L., and Brusic, V. (2011). MULTIPRED2: A computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles. *J. Immunol. Methods* 374, 53–61.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 August 2011; accepted: 14 November 2011; published online: 20 December 2011.

Citation: Olsen LR, Zhang GL, Keskin DB, Reinherz EL and Brusic V (2011) Conservation analysis of dengue virus T-cell epitope-based vaccine candidates using peptide block entropy. Front. Immun. ***2****:69. doi: 10.3389/fimmu.2011.00069*

This article was submitted to Frontiers in T-Cell Biology, a specialty of Frontiers in Immunology.