

DR. SANDEEP KUMAR DHANDA (Orcid ID : 0000-0003-1381-7434)

DR. ALBA GRIFONI (Orcid ID : 0000-0002-2209-5966)

Article type : Original Article

Development of a novel clustering tool for linear peptide sequences

Sandeep Kumar Dhanda^{1*}, Kerrie Vaughan¹, Veronique Schulten¹, Alba Grifoni¹, Daniela Weiskopf¹, John Sidney¹, Bjoern Peters^{1, 2}, Alessandro Sette^{1,2*}

Affiliations

¹ Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, California, United States of America

² Department of Medicine, University of California, San Diego, California, United States of America

* Corresponding authors

Alessandro Sette alex@lji.org , Sandeep Kumar Dhanda sdhanda@lji.org

Abstract

Epitopes identified in large-scale screens of overlapping peptides often share significant levels of sequence homology, complicating the analysis of epitope-related data. Clustering algorithms are often used to facilitate these analyses, but available methods are generally insufficient in their capacity to define biologically meaningful epitope clusters in the context of the immune response. To fulfill this need we developed an algorithm that generates epitope clusters based on representative or consensus sequences. This tool allows the user to cluster peptide sequences on the basis of a specified level of homology by selecting among three different method options. These include the 'clique method,' in which all members of the cluster must share the same minimal level of homology with each other, and the 'connected graph method,' in which all members of a cluster must share a defined level of homology with at least one other member of the cluster. In cases where it is not possible to define a clear consensus sequence with the connected graph method, a third option provides a novel 'cluster-breaking algorithm' for consensus sequence driven sub-clustering. Herein we demonstrate the tool's clustering performance and applicability using i) a selection of dengue virus epitopes for the 'clique method', ii) sets of allergen derived peptides from related species for the 'connected graph method' and iii) large datasets of eluted ligand, MHC binding and T cell recognition data captured within the Immune Epitope Database

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/imm.12984

This article is protected by copyright. All rights reserved.

(IEDB) with the newly developed 'cluster-breaking algorithm'. This novel clustering tool is accessible at <http://tools.iedb.org/cluster2/>

Introduction

Immune epitope data pertaining to specific antigenic regions often must be assessed from data generated using several different, although largely overlapping and/or highly homologous, sequences. This is typically because the exact sequences utilized to probe a specific region can differ between different laboratories, studies and methodologies, reflecting, for example, the use of different isolates, or alternative frame shifts.

Similarly, epitope identification strategies often involve the use of overlapping peptide sets, but can also include the testing of candidate peptides identified from MHC binding predictions^{1, 2}, as well as the elution of naturally-processed MHC ligands, which often occur as nested sequences³. Epitope-related sequences might also originate from homologous variants, such as different viral strains, homologous sequences found in related species, repeat sequences or single-nucleotide polymorphisms (SNPs)⁴. Different sequence variants might bind to closely related HLAs (e.g., HLA supertypes)⁵⁻⁷ and, in addition, certain sequences might reflect "hot spots" of immunodominant regions that are preferentially processed and bound⁸.

To capture this diversity, the Immune Epitope Database (IEDB) was specifically designed to curate each molecular structure (unique epitope) and associated metadata as a distinct entity⁹. However, in order to perform aggregate analyses on these individual data, there is a clear need for improved clustering strategies that facilitate biologically meaningful response analyses.

The IEDB has developed a tool, called Immunome Browser^{10, 11}, which enables visualization of the immune reactivity of related epitope sequences by displaying reported epitope reactivity along the linear sequence of a specified antigen or reference proteome. The tool is powerful for visualization purposes, but is not well suited to resolve redundancy of sequences in separate and specific clusters, and is also not ideal to analyze sequences that may cluster together, but are in fact derived from unrelated antigens. Indeed, the identification of regions of homology between unrelated sequences has been associated with a number of important biological phenomena, like molecular mimicry, allergen cross-reactivity, and microbiome modulation of adaptive immune responses¹²⁻¹⁶. Clustering epitope reactivity may also be of interest to generate epitope "megapools" of minimal composition by elimination of potentially redundant sequences based on specific degree of protein homology¹⁷⁻²⁰.

Clustering biological data, such as DNA or protein (amino acid) sequences, is a well-studied problem. However, making "biologically meaningful" clusters based on epitope response data presents unique challenges^{21, 22}. Here, a clustering approach would be biologically meaningful if it can provide connectivity of the entities in the clusters, provide a cluster representative sequence, and is based on sequence characteristics of each dataset (unsupervised strategy). Several attempts have been made in the past to address this issue from different angles. For example, Hammock²³ was developed to identify consensus motifs

in large datasets, and PepServe²⁴ to cluster peptides based on their physiochemical properties. Another application, UCLUST²⁵, performs rapid clustering and implements several previously identified clustering algorithms, and Andreatta et. al.^{26, 27} developed a tool based on Gibbs sampling. Overall, most of these tools cluster based on the physiochemical properties of peptide sequences or by defining and focusing on shared motifs.

However, none of these tools provide a complete connectivity of the peptides within a cluster and do not generate a clear consensus sequence representing each cluster. The need for elucidation of complete connectivity is key when evaluating cross-reactive responses or epitope homology between different species. Vice versa, generating a clear consensus sequence is important when trying to represent a group of sequences sharing a defined level of homology with a single unique sequence that maintains the same capacity to induce an immune response as the constituent epitopes, such as has been done with the megapool approach²⁰.

From a computational standpoint, two main clustering approaches exist: cliques²⁸ and clustering by connected graph²⁹⁻³¹. In the clique approach, sequences are clustered into cliques, defined as a group of sequences related to all other members at (or above) the exact desired level of homology. While this approach allows definition of clear consensus sequences, the main drawback is that the same sequence may be found in multiple cliques, leading to inflation of the number of clusters and ambiguity within clusters. That is, the clique approach provides greater inclusiveness, per se, but it also hinders cluster size reduction, which is desired for immune epitope clustering.

In the connected graph approach, all peptide sequences that are homologous to a certain pre-specified percentage level, for example 70%, are clustered together. In this case, any member of the cluster will be at least 70% homologous to at least one member of the cluster. While this approach clusters peptides only once, thereby avoiding redundancy, a drawback is that other members of the cluster might, and often are, related by levels of homology much lower than 70%, which is not suitable when you need to have a clear homology cutoff. In addition, this method cannot always resolve clearly consensus sequences.

To address these issues and apply them in different immunological contexts, we developed a clustering tool that is flexible enough to take into consideration different strategies, and generate consensus sequences from clusters by identifying the most commonly occurring residues at each position in a given length of amino acids based on a multiple sequence alignment. This novel tool allows the user to select different clustering algorithms and strategies depending on the biological question. We illustrate the performance of this method through the resolution of clusters of various epitope datasets available within the IEDB, and with different immunological purposes.

Methods

Dengue virus (DENV) dataset

The dengue virus epitope dataset was derived from a comprehensive analysis of HLA-matched CD4⁺ T cell responses in a Sri Lankan cohort, allowing for the selection of 363 epitopes restricted by 15 DRB1 alleles with phenotypic frequency higher than 4% in the Sri Lankan population, as previously reported¹⁸. The set of epitopes covers serotype-specific epitopes from DENV serotypes 1, 2, 3, 4 (all 10 proteins), as well as epitopes conserved between those serotypes. For each epitope a corresponding Spot Forming Colony (SFC) value is provided based on *in vitro* expansion of DENV-specific T cells and Interferon γ (IFN- γ) Enzyme-Linked Immunospot Assay (ELISA) to rank the best epitopes with highest SFC³².

Rat and mouse allergen datasets

Allergen epitope datasets were assembled from previous data characterizing rat and mouse antigens. The rat epitopes are derived from the rat allergen Rat n 1, a major urinary protein, and constitute a set of 19 peptides found to be T cell-reactive in rat allergic patients³³. In addition, our group recently identified 23 peptides from the major mouse allergen, *Mus m 1*³⁴.

IEDB datasets utilized for algorithm and tool development

Additional epitope datasets were compiled from the IEDB (<http://iedb.org>), which curates a vast set of published T cell response data, as well as MHC class I and II binding and ligand elution (MHCLE) data.

To retrieve relevant sets of epitopes, a query was performed from the IEDB home page targeting 'T cell Assays' (using check boxes in the 'Assay' search panel), and including both positive and negative peptides. As no other selection criteria were included in this initial query, these data represent both human and non-human data. The query was performed on February 2017, at which point the IEDB contained a total of 115,228 peptides tested (positive and negative) in T cell response assays. The full set of T cell response data was downloaded to Excel using the Assays tabs (Export T cell Assays Results).

To retrieve MHC binding and MHCLE data, a similar query was performed targeting (separately) each of these datasets (using 'Find' feature within the Assay search pane). These queries revealed a total of 64,312 peptides for MHC binding data, and 139,614 for MHCLE. From the exported Excel tables, we selected only linear peptides and further categorized each peptide dataset as associated with either class I or class II MHC molecules. Because of the unique immunobiology of the class I and class II, these datasets were analyzed separately in clustering algorithms. The composition and breakdown of the resulting dataset is summarized in S1 Table.

Generation of sequence homology matrices

In-house Python scripts were used to calculate the sequence homology between each peptide pair in each dataset. While calculating the homology between any peptide pair, one peptide is aligned to a second peptide in all the possible frames and the number of residues matching are counted for each frame (including the offsets). The alignment with the largest number of matches was used for homology calculations. To scale the level of sequence homology in the range of 0 to 1 (meaning from 0% to 100%), we divided the maximum number of matches in the alignment to the length of smaller peptide (see equation i).

$$\text{Sequence identity} = \frac{\text{Maximum number of residue matching in an alignment}}{\text{Length of smaller peptide}} \quad (\text{Equation i})$$

Building a graph network connecting clusters and cliques

The sequence homology matrix calculated as above was then used to create an undirected graph network. Here, an undirected graph network works as a spacial/graphical representation of the data, where nodes within the network represent the clustered peptides and the edges represent the sequence homology defined above for a given threshold. The undirected graph network was generated with the Networkx version 1.11 library of Python language³⁵. Using this, all the fully interconnected nodes (called cliques) and the connected components (called a connected graph) were derived. Here, a clique is defined as a set of peptides, wherein all the peptides share sequence homology higher than the threshold with all other members of the clique. By contrast, each connected graph represents a group of peptides, where each peptide shares sequence homology higher than the threshold to at least one of the other peptides in that connected graph. As a result, one peptide can be present in more than one clique, whereas each cluster contains only unique peptides, and no peptide was present in two connected graphs.

Generation of alignments and consensus sequences from peptide data sets

Sequence alignments were generated using Clustal Omega version 1.2.4 for high quality alignments based on several benchmark studies³⁶. From the alignment output, the consensus sequence is derived using the Biopython package³⁷. Here, we used a threshold of 51% occurrences to assign a residue at a specific position. An X was assigned for ambiguous residues. These alignments were used to derive consensus sequences from each set of peptides in a cluster.

Cluster breaking algorithm applied on connected graph method and web implementation

A cluster breakdown algorithm was developed scanning each connected graph for the longest possible consensus sequence using an undirected graph network constructed from a set of peptides (as described above). The process starts by extracting the most connected peptide (center node, N0) in a graph, including all the directly connected peptides (neighbors, N1) of that peptide. Then the next set of peptides (N2) is extracted, and connected to center node (N0) through N1. Each peptide in N2 is then aligned with a consensus sequence, and peptides that share sequence homology less than the threshold of the consensus sequence are removed. The next alignment is created with the remaining peptides from N2 and all the peptides from N1 (including center node). A new consensus is then derived from this alignment. The process is iterated until the clear longest consensus sequence that can represent the maximum number of peptides in a connected graph is obtained. These peptides are then removed from the network and the process is repeated in a recursive fashion for each sub-graph. The complete cluster-break pipeline was implemented in the Django framework and can be accessed by users as an online tool. Javascripts were used to support the visualization in the server.

Results

Use of a network graph approach to generate cliques and connected graph

Our overall goal was to develop an online tool that would allow the user to apply different clustering strategies to a list of epitope or peptide sequences. Specifically, we used two main methodologies to cluster sequences, one by cliques²⁸ and the other by connected graph²⁹⁻³¹.

For this we developed a simple algorithm to cluster sequences using a network graph approach³⁵. A network was created using a matrix, where each row represents a peptide pair and their sequence homology. This method creates graphical networks of sequences in which all members of the cluster are homologous to all other members of the connected graph at a specified homology level (S1 Fig). Here circles, or nodes, represent the peptides and the solid line is the homology above threshold between any peptide pair. From these graphical networks, peptide sequences that are connected in the network (called as connected graph) can be extracted. The peptides labeled as A, B, C and D are connected graphs in S1 Fig. As an alternative, all the peptides that are fully interconnected (referred to as cliques) can be extracted. The cliques are represented by peptides A, B and C (clique 1) or peptides B, C and D (clique 2) in the example above, as peptides A and D are not connected (S1 Fig). Peptides (e.g., E in our example) which do not share any homology with remaining peptides, are referred to as singletons. Each peptide will be present in only one connected graph, but can be present in multiple cliques. Singletons are represented the same way in the connected graph and clique approaches.

Application of the clique method to refine a previously identified DENV-derived epitope megapool

Next, we evaluated the performance of our clustering methods using different approaches suited to the different immunological questions behind it. As a first example we used a set of 363 DENV derived epitopes for which a “megapool” of peptides was previously generated based on experimental data from HLA-matched CD4⁺ T cell responses in a Sri Lanka DENV-endemic population¹⁸. In this case, we sought to reduce the number of peptides included in the megapool to facilitate experimental approaches without sacrificing biological activity. For this purpose, to obtain clear consensus sequences encompassing peptides achieving the same homology level cutoff despite the possibility of having cluster redundancy, the clique method was applied.

The results are summarized in Table 1, and also enclosed in the Tables S2-S3. Among the original 363 DENV epitopes, 305 were unique in the dataset and at 70% sequence identity they were grouped into 198 cliques (Table S2). Of those 198 cliques, 118 cliques contained two or more epitopes and, in total, covered 225 epitopes. The remaining 80 epitopes did not share any homology with the rest of the epitope set and hence were assigned as singletons (clusters containing a single sequence).

Examining the best epitopes (i.e., those with highest magnitude of response, in terms of SFCs) contained in each of the 118 cliques, we observed that this corresponded to 83 unique epitopes (some epitopes were selected as the best epitope in more than one clique). Thus, in the end, the clique clustering generated a set of 80 singletons and 83 unique “best of the clique” epitopes, or a total of 163 epitopes that could be utilized to represent the 363 DENV “megapool” previously identified, equivalent to a 55% reduction in the number of epitopes needed (Table S3).

Application of the clustering method in mouse versus rat allergen data

As another demonstration of the peptide-clustering application we analyzed data derived from a study of mouse allergens. To date, one major mouse allergen, Mus m 1, has been identified, to which our lab has mapped 23 different T cell epitopes³⁴. A previous study had identified 19 T cell epitopes³³ from the major rat allergen, Rat n 1, which exhibits significant homology to Mus m 1 (65% homology, 78% similarity), making these datasets good candidates for cluster analysis.

Here the purpose of the clustering task was to judge whether each epitope derived from the mouse allergen would co-cluster with sequences derived from the rat allergen. For this, the clique approach was not appropriate, since the evaluation of the association requires that each sequence be represented in only one cluster or it will inflate the amount of sequences shared between the two species. Accordingly, the connected graph strategy was used (Table 2).

The analysis revealed that the 23 overlapping Mus m 1 epitopes correspond to 20 unique epitopes, as three epitopes clustered with other epitopes within the same set (70%

homology threshold). Of the 23 Mus m 1 epitopes, 8 clustered with Rat n 1 peptides. Six of the eight Mus m 1 epitopes that clustered with Rat n 1 peptides are among the top nine in terms of T cell response magnitude (conversely, only 2 peptides are in the bottom fourteen; $p=0.023$ by exact Fisher test). This analysis supported the notion that epitope allergens conserved in multiple species may elicit a more dominant immune response, likely due to increased/repeated exposure by the various homologous allergens.

Analysis of MHCLE data reported in the IEDB

We then tested the two clustering approaches on a larger dataset consisting of naturally-processed peptides eluted from MHC, using MHC ligand elution (MHCLE) data reported in the IEDB. MHCLE peptides are of variable size, and often the sequences eluted are largely nested and/or overlapping^{38, 39}. For this reason, this type of data was ideal for further validation of our clustering tool and test the applicability on a larger scale data.

The IEDB has MHCLE data for over 100,000 different peptide sequences. Sets of class I and class II naturally processed peptides were generated by querying the IEDB (as described in the methods) and the extracted sequences were clustered using the network graph approach³⁵. As an alternative, all the peptides that were fully interconnected (cliques) were also extracted.

Because of their unique immunobiology, class I and II ligands were analyzed separately. We expected that the effect of sequence clustering would be more pronounced for class II, because of the well-known “ragged end” nature of natural class II ligands⁴⁰, allowing ligands of variable length. Indeed, we found that of a total of 105,642 class I and 33,757 class II peptides, 57,455 and 28,523 (respectively) were clustered (Table 3), and 48,187 and 5,234 were singletons (not clustering with any other sequence) for class I and class II, respectively. Thus, almost half (46%) of the class I peptides were identified as singletons compared to only 16% for class II; 84% of the class II sequences were found in clusters (Table 3).

Using the connected graph (where different clusters can be connected through single peptide, Fig S1) clustering method, we found 11,932 and 4,683 clusters, respectively, for class I and II. The average cluster sizes were 4.82 and 6.09 peptides, respectively, indicating that this method can effectively cluster the data. By contrast, we observed 36,732 and 25,941 cliques, and the average clique sizes were 2.78 and 57.36 for the two datasets in clique method. Thus, there are many more cliques than clusters in class II and the average size of cliques (57.36 peptides, on average) is much greater than the average size of clusters (6.09 peptides, on average). This is due to the fact that many peptides are found in multiple cliques, and indicates that the clique approach was ineffective for dataset reduction. As a result, it was not considered further for this particular application.

We then applied the clustering method on the same larger datasets evaluating the distribution of connected graph clusters generated with the class II MHCLE data as a function of the number of peptides contained in each cluster. The distribution is shown in Fig 1 (red line) for the top ten clusters (solid red line) and also summarized in Table 4 (“Raw Data” column). Table 4 shows that a single cluster contained more than 1000 peptides and a total of 10 clusters contained between 100 – 999 peptides. Similarly, we observed 21

clusters with 50-99 peptides, 55 clusters with 30-49 peptides, 439 clusters with 10-29 peptides. Finally, the majority of the clusters (4,157 out of 4,683:(89%) encompassed less than 10 peptides.

However, manual inspection of the “large membership” clusters (i.e., clusters with >1000 peptides) revealed that they contained rather heterogeneous sequences with very little actual sequence similarity, often linked by the presence of very short sequence stretches of 3-4 residues. For several of the large clusters a clear consensus sequence could not be defined (data not shown). We reasoned that the largest and smallest sequences contained in the data pollute the clustering assignments, and thus sought to alleviate this problem by instituting appropriate peptide length/size filters, and developing an additional feature on the clustering method.

Resolution is improved by excluding tail-end sequence sizes

To implement an appropriate filter system for peptide length, we reasoned that the large majority of biologically relevant data (MHC ligands, MHC binding or T cell recognition) are associated with epitopes between 8-25 residues⁴¹⁻⁴⁴. To test this assumption, we generated datasets corresponding to class I and class II peptides in the MHLE, MHC binding and T cell response data.

Across the different datasets, there were few peptides with length less than 8 residues (<8mer) (Table 5). The range for <8mer was from as low as 0.09% (T cell assay for MHC class I) to 0.63% (MHCLE class I). The majority of the peptides for MHC class I (~98%) were represented by peptides 8-15aa in length for the MHCLE and MHC binding datasets, and 8-20aa for peptides within the T cell assays dataset. To achieve 98% coverage for class II, the length range is higher; we consider peptides 8-25aa in length for MHCLE and 9-21aa for the MHC binding datasets. When we examined the size distribution of all these various ligand categories, we found that filtering the peptide datasets on the 8-25aa size range captured 98% or more of all ligands, except for T cell assays for MHC class II, where it covers up to 97.24% of the peptides. Accordingly, in subsequent analyses, we only considered peptides in the 8-25 residue range, thus effectively removing outlier sequences from the top and bottom end of the length distribution for both class I and II (Table 5).

Next, we repeated the cluster analysis considering only peptides in the 8-25 residue range. As expected, several of the large clusters were broken down into smaller ones, and the number of clusters increased from 4,683 to 4,726. Maximum cluster size dropped from 1,006 to 649, with only 11 clusters containing more than 100 peptides (Table 4; “Length filtered column”). This normalizing effect was also observed in the plot representing the number of peptides in each cluster (Fig 1 Blue line). Also, as expected, the average number of peptides in each of the different clusters was decreased. In most cases a clear consensus sequence could be defined, but for several of the large clusters a clear consensus sequence remained elusive (data not shown).

In conclusion, while the data filtering increased the clustering resolution, the presence of several large and ambiguous clusters suggested that further approaches were required to maximize clustering resolution in the case of large datasets like the ones considered herein.

Development of an algorithm that allows cluster-break for clear separation

Manual inspection of the remaining large clusters revealed that in several cases, these clusters contained numerous loosely connected sub-clusters (Figure 2A). To visualize this feature, we followed the approach of constructing an undirected graph network from peptide sequence homology matrices using a publicly available network package³⁵. According to this approach, a circle represents each peptide sequence, and any peptides (circles) that share an indicated level of homology (in this case 70%) are connected by a solid line (Fig S1).

Application of the cluster-break algorithm to various large peptide datasets

When the cluster-break algorithm was applied to the MHC class II ligand elution dataset, the number of clusters increased to 4,912, and the largest clusters now contained only 219 sequences (Table 4; “Length filtered” and “cluster break algorithm” columns). The largest sub-cluster was a “legitimate” cluster with the consensus sequence “AASQRMEPRAPWIEQEGPEYWDXETRXVKAHSQTH”. A further smoothing in the distribution of peptides in different clusters is also observed (Fig 1 green line). Based on these results we surmise that the combination of the peptide size filter and the “cluster-breaking” algorithm represents a viable clustering strategy for large datasets of this type.

An example of this representation based on a cluster composed of 99 different peptides is shown in Fig 2A. Next, a cluster-break algorithm was derived to further dissect clusters into sub-clusters, and achieve definition of the longest consensus sequence (Fig 2B), as described in the methods.

To further validate the cluster-break strategy, we performed a similar analysis on the class I MHCLE dataset, and MHC binding and T cell epitopes for both class I and class II datasets. We observed no cluster with more than 1,000 members (Table 6). The maximum size for any cluster was 257, found in one case in the MHC class I binding dataset (Fig 3, dashed blue line). The maximum cluster size for class I peptides in the MHCLE and TCR datasets was in the range of 50-99, whereas the range was 100-999 for the remaining datasets. Similar to the MHCLE class II dataset, there were few clusters in the size range of 30-49 or 10-29, whereas the majority of clusters have less than 10 peptides. Fig 3 plots the top 10 clusters from each dataset after applying the cluster-break algorithms.

Further analysis of the overlap between MHC binding, MHC Ligand elution (MHCLE) and T-cell data

We then focused on a different immunological question that can be addressed analyzing large datasets. In this case we compared MHC binding data with MHCLE and T cell response in order to identify epitopes that were redundant in the three datasets and could therefore be the best candidates able to be presented and recognized by T cell.

For this purpose, we clustered all the T cell epitopes, MHC binding and MHCLE data in pairs to analyze the homology sharing between MHC binders or MHC eluted data with T cell response. As above, class I and class II data were analyzed separately. Once we obtained the results of the clustering we further categorized each cluster as positive for T cell assays if it contained at least one sequence associated with positive T cell assay results, and as negative if it contained only sequences negative in T cell assays. A cluster was classified as T cell assay undetermined if none of its sequences was associated with either positive or negative T cell recognition. A similar classification was performed to classify clusters according to MHC binding assay status. Unlike MHC binding and T cell epitope data, which represent both positive and negative assays, for MHCLE data there are only positive data (MHCLEp) (S4 Table), therefore it cannot be inferred whether the sequence is negative or not tested.

In the case of class I, out of a total of 2,719 T cell assay + clusters, only 515 were undetermined (19%); of the remaining clusters, 2,026 were MHC+ (74.5%) (Fig 4A). Similarly, in the case of class II, 1260 (38.6%) of the T cell assays + clusters were also MHC+ (Fig 4B). Conversely, when we examined MHC+ clusters we found that 28.6% and 47.6% were also T cell assay +, for class I and class II, respectively. Taken together, this analysis supports the notion that MHC binding is necessary, but not sufficient, condition for T cell immunogenicity.

When we examined MHCLE data, we found that the vast majority of MHCLEp clusters are undetermined when it comes to T cell assays (85.1 % and 89.9% for class I and class II respectively) (Fig 4C-4D) or MHC binding (83.5 % and 91.5% for class I and class II respectively) (Fig 4E-4F). In the few cases where the cluster could be categorized, the results were consistent with expectations, in that the large majority (82.4% for class I to 87.6% for class II) of MHCLEp clusters were also positive for MHC binding assays. We expected that being positive for MHCLE would have been a better predictor of T cell assay positivity than simple MHC binding. Surprisingly, this was not the case and we found that 42% and 77.8% MHCLEp clusters were also T cell assay +, for class I and class II, respectively. This is in agreement with the notion that MHC binding and ligand processing are necessary, but not sufficient for T cell immunogenicity⁴⁵.

Implementation of the three methods in a unique clustering tool online

An online version of the tool described herein was developed, with the aim of making it accessible through the IEDB website. We designed the tool to be as user-friendly as possible by requiring minimum parameter input from the users, but at the same time preserving rigorous functionality.

In terms of an input interface, users will provide a set of sequences by pasting directly into a text box or through upload in a given file format (Fig 5A). The sequences can be in either plain text format, where each line has a separate sequence, or in FASTA format. In addition, the user can attach metadata to the sequences, such as an IEDB peptide identifier or other user-generated metadata (such as the antigen or organism source of the sequence, sequence position etc.) in the FASTA formatted file. Next, the user chooses the minimum threshold for sequence homology, and whether to apply the size-exclusion filter (default, no length filters) (Fig 5B). And then, in a third step, the user would choose one of three available options for the clustering approach (Fig 5C). These three options correspond to the connected graphs, cluster-break and cliques approach. In default settings, cluster-break algorithms will be run.

Once the user has filled in these options, a submit button will automatically redirect the user to the results page. The results page displays a summary of results, including the total number of peptides, unique number of peptides, selected homology threshold and total number of clusters found in given dataset (Fig 5D). The option of exporting the result to Excel will be provided.

The complete results are provided both in tabular and graphic form. The tabular results show the consensus or peptide number, peptide alignment, position of a peptide in the alignment, description and the peptide sequence against each cluster number. The cluster number is represented by a decimal number, where the number before the decimal point is the cluster number before applying the cluster-breaking algorithm and the number following the decimal point identifies each sub-cluster identified by the algorithm.

In the graphic representation (Fig 5E), each cluster is visualized in the form of networks, where each node is a peptide and the edge is the connection between any two nodes, if they have homology greater than the selected threshold. Single nodes correspond to peptides where no other sequence shares more than the given threshold of homology.

In the current version of clustering tool (Cluster2), we are offering the three different types of approaches here described, which are useful in different contexts of peptide application as summarized in Table 7. The clique approach generates clusters of fully interconnected peptides according to the specified identity threshold. The approach will give, for each cluster, a clear representative sequence, but cluster redundancy would be expected. This approach is useful in generating, for example, megapools where all the peptides belonging to a single cluster have the same sequence identity threshold, and it would be efficacious to select one peptide (based on user discretions, in our case higher magnitude of response in terms of SFCs) from each cluster. Unique peptides can be selected at the end to address redundancy issues.

The cluster approach groups all directly or indirectly connected peptides, based on specified identity threshold. In this case, peptides belonging to the same cluster might have a sequence identity below the threshold, but no peptide redundancy is observed. This approach would be useful, for example, to create non-redundant clusters of peptides which can be used to analyze cross-reactivity with a minimal peptide set. For larger datasets, representative sequence/s might not be clear.

The cluster-break approach is an extension of above mentioned cluster approach, where clusters are broken down to sub-clusters to obtain a clear representative sequence from each sub-clusters and to address issues of large membership. Therefore, this approach is very useful in applications where it would be helpful to reduce the number of peptides derived from larger datasets.

Comparison with the existing algorithms

Clustering is a well-studied problem in biological data, and has been applied to address different biological questions like creating gene networks, evolutionary relationships and to find sequence homologs. Due to the different purposes for which various algorithms have been developed, it is somewhat challenging to compare an algorithm with others based on a benchmarking dataset. Instead, here, we compared the features of different algorithms for peptide clustering (Table 8). In addition to the approaches described herein, we considered, for feature comparison, GibbsCluster^{26, 27}, PepServe²⁴, Hammock²³ and UCLUST²⁵. Table 8 provides a summary of the various features assessed in the ensuing.

Some critically important features are taken into consideration for comparison, which include user input sequences, free online tool, graphical visualization, connectivity of peptides within cluster, cluster representative sequence, dealing with large datasets, basis of calculating identities among peptides and unsupervised nature of clustering (Table 8). As evident from the comparison, the new version of clustering tool (cluster2) offers many missing features over existing tools. Overhang identity calculations, visualizing connectivity and providing clear representative sequences are some unique features implemented in the tool. Additionally, our novel cluster-breaking algorithm provides unique way of clustering in large datasets.

Discussion

Peptide clustering has several potential applications in an immunological context. While there are various algorithms available for data clustering⁴⁶⁻⁴⁸, the identification of regions with specific homology, as well as representation of those with clear consensus sequences, has been challenging for the development of appropriate clustering algorithms.

In the past, several methods have been published investigating different aspects of peptide clustering, such as UCLUST²⁵, PepServe²⁴, Gibbs clustering^{26, 27} and the Hammock algorithm²³. However, all these tools do not provide the complete connectivity of the peptides within a cluster and a biologically meaningful consensus sequence representing each cluster, features that are of great utility, particularly in the context of immunological studies. In addition, as many available clustering tools use a specific approach, it is challenging for immunologists to select the appropriate algorithms able to address the biological question of interest.

Here, we extend graph theory applications to peptide clustering, where connectivity is defined by shared sequence identity cut-off. Accordingly, peptides that are fully homologous to each other are grouped in cliques, whereas peptides that are homologous to a point are grouped together in connected components. Both of these approaches have their own pros and cons. In the clique approach, each group can give a clear representative sequence, but a peptide can be present in several groups, hindering dataset reduction. In the peptide-connected approach, no peptide would be present in two groups, but the group may not provide a clear representative sequence. In addition, we report an alternative approach, where no peptide would be present in more than one group, and also each group can provide a clear representative sequence. In this approach, we develop a novel cluster-break technique allowing derivation of a clear consensus sequence.

The three different approaches are designed to address different biological questions. The clique approach selects a clear consensus sequence out of each clique and have all the peptides grouped on a specific homology level cutoff. This was applied on DENV data, and can be extended in all the cases where redundancy is not important, while a precise identity level is required in each cluster. The reason for applying the clique approach in general is that the peptides in each clique are fully interconnected and selecting the peptide with highest response can represent the remaining peptides in that clique. Thus, the clique approach is suitable for highly redundant data with sequence overlap and or high sequence homology. A similar clique-based approach has previously been applied to analyze the glycan structures⁴⁹ and biological data mining⁵⁰.

Alternatively, the cluster approach based on connected components, contains sequences with a lower level of homology respect to the one used as cutoff, but where no peptide redundancy is observed in each cluster. The cluster approach produced significant and meaningful clusters for T cell epitopes derived from rat and mouse allergens³⁴. In this instance, the resulting clusters shared high interconnectivity and were smaller in size, thus a clear representative/consensus sequence was observed and similarities among the two species were not inflated by the presence of redundancies.

Finally, in the context of larger datasets, we observed that neither the clique approach nor the connected components provide legitimate clusters. For this reason, we generated an algorithm to break down clusters in a systematic manner to extract a clear representative/consensus sequence from each sub-cluster. This cluster-break approach has shown to be the most suitable in the context of variable and larger datasets such as in the case of known epitopes derived from the IEDB⁹. This approach is appropriate for applications where the user, for example, would want to derive minimal, non-redundant, and clear representation from large datasets or identify common sequences between different datasets as in the case of epitopes able to both bind to MHC and induce T cell response. It is worth noticing that comparable efficacy for both CD4+ and CD8+ T cells clustering analysis is observed in the context of IEDB- large dataset example for the cluster-break method, suggesting that the tool shows equal efficacy in both the type of datasets.

Herein we describe the creation of a freely available online tool for peptide clustering. The tool provides a user-friendly format to support clustering of disparate epitope datasets. The unique tool incorporates multiple approaches to analyze different types of dataset and address different immunological questions. We applied the tool to in house experimentally derived CD4+ restricted epitopes, however the tool is equally applicable to CD8+ T cell

epitopes, as shown the IEDB-derived dataset. In addition to the epitopes, the tool can also be applied to cluster and explore interesting patterns in any linear peptide dataset (like Cell penetrating peptides^{51, 52}, tumor homing peptides⁵³, hemolytic peptides⁵⁴, therapeutic peptides etc⁵⁵.)

Acknowledgments

Authors acknowledge the funding supported by U.S. National Institutes of Health grants HHSN27220140045C, HHSN272201200010C and 1P01AI106695-01A1 to A.S.

Author's contribution

SKD wrote the code, analyzed the data with AS and BP. VS contributed the rat and mouse allergy section. KV provided the larger dataset. AG and DW analyzed the DENV data. AS and BP conceived and supervised the project. JS edited the manuscript. All authors approved the manuscript.

Conflict of interest

Authors declare no competing financial interests.

References

1. Vivona S, Gardy JL, Ramachandran S, Brinkman FS, Raghava GP, Flower DR, et al. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends Biotechnol* 2008; 26:190-200.
2. Dhanda SK, Usmani SS, Agrawal P, Nagpal G, Gautam A, Raghava GPS. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. *Brief Bioinform* 2017; 18:467-78.
3. Paradela A, Alvarez I, Garcia-Peydro M, Sesma L, Ramos M, Vazquez J, et al. Limited diversity of peptides related to an alloreactive T cell epitope in the HLA-B27-bound peptide repertoire results from restrictions at multiple steps along the processing-loading pathway. *Journal of Immunology* 2000; 164:329-37.
4. Altenburg AF, Rimmelzwaan GF, de Vries RD. Virus-specific T cells as correlate of (cross-)protective immunity against influenza. *Vaccine* 2015; 33:500-6.
5. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 1999; 50:201-12.
6. Fikes JD, Sette A. Design of multi-epitope, analogue-based cancer vaccines. *Expert Opin Biol*

- Ther 2003; 3:985-93.
7. Ou D, Mitchell LA, Tingle AJ. HLA-DR restrictive supertypes dominate promiscuous T cell recognition: association of multiple HLA-DR molecules with susceptibility to autoimmune diseases. *J Rheumatol* 1997; 24:253-61.
 8. Ayyoub M, Merlo A, Hesdorffer CS, Speiser D, Rimoldi D, Cerottini JC, et al. Distinct but overlapping T helper epitopes in the 37-58 region of SSX-2. *Clinical Immunology* 2005; 114:70-8.
 9. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015; 43:D405-12.
 10. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Front Immunol* 2017; 8:278.
 11. Dhanda SK, Vita R, Ha B, Grifoni A, Peters B, Sette A. ImmunomeBrowser: A tool to aggregate and visualize complex and heterogeneous epitopes in reference protein. *Bioinformatics* 2018.
 12. Mukherjee S, Warwicker J, Chandra N. Deciphering complex patterns of class-I HLA-peptide cross-reactivity via hierarchical grouping. *Immunol Cell Biol* 2015; 93:522-32.
 13. Stufano A, Capone G, Pesetti B, Polimeno L, Kanduc D. Clustering of rare peptide segments in the HCV immunome. *Self Nonself* 2010; 1:154-62.
 14. Lucchese G, Capone G, Kanduc D. Peptide sharing between influenza A H1N1 hemagglutinin and human axon guidance proteins. *Schizophr Bull* 2014; 40:362-75.
 15. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 2017; 547:89-93.
 16. Backert L, Johannes Kowalewski D, Walz S, Schuster H, Berlin C, Christoph Neidert M, et al. A meta-analysis of HLA peptidome composition in different hematological entities: Entity-specific dividing lines and pan-leukemia antigens. *Oncotarget* 2017.
 17. Lindestam Arlehamn CS, McKinney DM, Carpenter C, Paul S, Rozot V, Makgohlho E, et al. A Quantitative Analysis of Complexity of Human Pathogen-Specific CD4 T Cell Responses in Healthy M. tuberculosis Infected South Africans. *PLoS Pathog* 2016; 12:e1005760.
 18. Weiskopf D, Angelo M, Zapardiel J, Seumois G, de Silva A, de Silva AD, et al. DENV-specific CD4 T-cells dominantly recognize capsid-derived epitopes and display a cytotoxic phenotype. *Journal of Immunology* 2016; 196.
 19. Hinz D, Oseroff C, Pham J, Sidney J, Peters B, Sette A. Definition of a pool of epitopes that recapitulates the T cell reactivity against major house dust mite allergens. *Clin Exp Allergy* 2015; 45:1601-12.
 20. Carrasco Pro S, Sidney J, Paul S, Lindestam Arlehamn C, Weiskopf D, Peters B, et al.

- Automatic Generation of Validated Specific Epitope Sets. *J Immunol Res* 2015; 2015:763461.
21. Zhao Y, Karypis G. Clustering in life sciences. *Methods Mol Biol* 2003; 224:183-218.
 22. Rottger R. Clustering of Biological Datasets in the Era of Big Data. *J Integr Bioinform* 2016; 13:300.
 23. Krejci A, Hupp TR, Lexa M, Vojtesek B, Muller P. Hammock: a hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinformatics* 2016; 32:9-16.
 24. Alexandridou A, Dovrolis N, Tsangaris GT, Nikita K, Spyrou G. PepServe: a web server for peptide analysis, clustering and visualization. *Nucleic Acids Research* 2011; 39:W381-W4.
 25. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010; 26:2460-1.
 26. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research* 2017; 45:W458-W63.
 27. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 2013; 29:8-14.
 28. Peay ER. Hierarchical Clique Structures. *Sociometry* 1974; 37:54-65.
 29. Gibbons A. Algorithmic graph theory. Cambridge Cambridgeshire ; New York: Cambridge University Press; 1985.
 30. Bondy JA, Murty USR. Graph Theory with Applications: North Holland; 1976.
 31. West DB. Introduction to Graph Theory: Prentice Hall; 2008.
 32. Weiskopf D, Angelo MA, Grifoni A, O'Rourke PH, Sidney J, Paul S, et al. HLA-DRB1 Alleles Are Associated With Different Magnitudes of Dengue Virus-Specific CD4+ T-Cell Responses. *J Infect Dis* 2016; 214:1117-24.
 33. Jeal H, Draper A, Harris J, Taylor AN, Cullinan P, Jones M. Determination of the T cell epitopes of the lipocalin allergen, Rat n 1. *Clin Exp Allergy* 2004; 34:1919-25.
 34. Schulten V, Westernberg L, Birrueta G, Sidney J, Paul S, Busse P, et al. Allergen and Epitope Targets of Mouse-Specific T Cell Responses in Allergy and Asthma. *Front Immunol* 2018; 9:235.
 35. Aric AH, Daniel AS, Pieter JS. Exploring network structure, dynamics, and function using NetworkX. 2008:11--5.
 36. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 2011; 7.
 37. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics.

- Bioinformatics 2009; 25:1422-3.
38. Sette A. Naturally Processed Peptides; 1993.
39. Schellens IMM, Hoof I, Meiring HD, Spijkers SNM, Poelen MCM, van Gaans-van den Brink JAM, et al. Comprehensive Analysis of the Naturally Processed Peptide Repertoire: Differences between HLA-A and B in the Immunopeptidome. Plos One 2015; 10.
40. Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K, et al. Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. Science 1992; 256:1817-20.
41. Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DA, et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature 1992; 358:764-8.
42. Sercarz EE, Maverakis E. MHC-guided processing: Binding of large antigen fragments. Nature Reviews Immunology 2003; 3:621-9.
43. Chang ST, Ghosh D, Kirschner DE, Linderman JJ. Peptide length-based prediction of peptide-MHC class II binding. Bioinformatics 2006; 22:2761-7.
44. Rist MJ, Theodossis A, Croft NP, Neller MA, Welland A, Chen Z, et al. HLA peptide length preferences control CD8+ T cell responses. J Immunol 2013; 191:561-71.
45. Dhanda SK, Karosiene E, Edwards L, Grifoni A, Paul S, Andreatta M, et al. Predicting HLA CD4 immunogenicity in human populations. Frontiers in Immunology 2018; 9:1369.
46. Aggarwal CC, Reddy CK. Data Clustering: Algorithms and Applications: Taylor & Francis; 2013.
47. Xu R, Wunsch DC, 2nd. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng 2010; 3:120-54.
48. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 2009; 10:297-314.
49. Fukagawa D, Tamura T, Takasu A, Tomita E, Akutsu T. A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures. BMC Bioinformatics 2011; 12 Suppl 1:S13.
50. Matsunaga T, Yonemori C, Tomita E, Muramatsu M. Clique-based data mining for related genes in a biomedical database. BMC Bioinformatics 2009; 10:205.
51. Agrawal P, Bhalla S, Usmani SS, Singh S, Chaudhary K, Raghava GP, et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res 2016; 44:D1098-103.
52. Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, et al. CPPsite: a curated database of cell penetrating peptides. Database (Oxford) 2012; 2012:bas015.
53. Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GP. TumorHoPe: a database of tumor homing peptides. PLoS One 2012; 7:e35187.

- Accepted Article
54. Gautam A, Chaudhary K, Singh S, Joshi A, Anand P, Tuknait A, et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res* 2014; 42:D444-9.
 55. Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, et al. SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res* 2015; 44:D1119--D26.

Table 1: Summary of the DENV data clustering using clique approach

Feature	Counts
Total peptides	363
Unique peptides	305
Total cliques	198
Peptides covered in cliques with ≥ 2 peptides	225
Cliques with ≥ 2 peptides	118
Unique peptides selected in cliques with ≥ 2 peptides	83
Singletons cliques	80
Total peptides selected	163

Table 2: The clustering tool output with mouse and rat allergic data.

Cluster Number	Peptide Number	Alignment	Position	Description	Peptide
1	Consensus	TFQLMXLYGRXXDLSSDIKEKFAKLCEA	-	-	-
1	1	TFQLMVLYGRTKDLSSDIKE-----	1	Rat Pep17	TFQLMVLYGRTKDLSSDIKE
1	2	-----GLYGREPDLSDDIKERFA-----	6	Mus Pep3	GLYGREPDLSDDIKERFA
1	3	-----YGREPDLSLDIKEK-----	8	Mus Pep7	YGREPDLSLDIKEK
1	4	-----GRTKDLSSDIKEKFAKLCEA	9	Rat Pep9	GRTKDLSSDIKEKFAKLCEA
2	Consensus	YDRYVMXHLINXXGETFQLMXLYGR TK	-	-	-
2	1	YDRYVMFHLINFKNGETFQL-----	1	Rat Pep19	YDRYVMFHLINFKNGETFQL
2	2	-----AHLIN EKDG E T F Q L M-----	7	Mus Pep9	AHLIN EKDG E T F Q L M
2	3	-----LINFKNGETFQLMVLYGR TK	9	Rat Pep12	LINFKNGETFQLMVLYGR TK
2	4	-----NEKDG E T F Q L M G L Y-----	11	Mus Pep6	NEKDG E T F Q L M G L Y
3	Consensus	EENGSMRVFXXHIXVLENSL	-	-	-
3	1	EENGSMRVFMQHIDVLENSL	1	Rat Pep4	EENGSMRVFMQHIDVLENSL
3	2	---GSMRVFVEHIHVLEN--	4	Mus Pep16	GSMRVFVEHIHVLEN
4	Consensus	FXXHIXVLENSLXFKFRIKE	-	-	-
4	1	FMQHIDVLENSLGFKFRIKE	1	Rat Pep6	FMQHIDVLENSLGFKFRIKE
4	2	FVEHIHVLENSLAFK-----	1	Mus Pep2	FVEHIHVLENSLAFK
5	Consensus	RXNIIDLTKTXRCLXARG	-	-	-
5	1	RDNIIDLTKTDRLQARG	1	Rat Pep14	RDNIIDLTKTDRLQARG
5	2	-ENIIDLTKTNRCLKA--	2	Mus Pep17	ENIIDLTKTNRCLKA
6	Consensus	GXWFSIXXASXKREKIEENG	-	-	-
6	1	GDWFSIVVASNKREKIEENG	1	Rat Pep8	GDWFSIVVASNKREKIEENG
6	2	-EWF SILLASDKREKI----	2	Mus Pep4	EWF SILLASDKREKI
7	Consensus	EEASSTGRNFNVXKINGEWHTIIL	-	-	-
7	1	EEASSTGRNFNVQKINGEWHTIIL	1	Mus Pep10	EEASSTGRNFNVQKINGEWHTIIL
7	2	-----NVEKINGEWHTIIL	11	Mus Pep13	NVEKINGEWHTIIL
8	Consensus	FVEYDGXNTFTILKTDYDXY	-	-	-
8	1	FVEYDGGNTFTILKTDYDRY	1	Rat Pep7	FVEYDGGNTFTILKTDYDRY
8	2	----DGNTFTILKTDYDN-	5	Mus Pep5	DGNTFTILKTDYDN
9	Singleton	TFTILKTDYDRYVMFHLINF	-	Rat Pep18	TFTILKTDYDRYVMFHLINF
10	Singleton	GIYYLNYDGFNTFTI	-	Mus Pep14	GIYYLNYDGFNTFTI

11	Singleton	KTPEDGEYFVEYDGGNTFTI	-	Rat Pep10	KTPEDGEYFVEYDGGNTFTI
12	Singleton	LENSLVLFHTVRDE	-	Mus Pep8	LENSLVLFHTVRDE
13	Singleton	LQSGFYSLSSLVTP	-	Mus Pep21	LQSGFYSLSSLVTP
14	Singleton	ENSLGFKFRIKENGECRELY	-	Rat Pep5	ENSLGFKFRIKENGECRELY
15	Singleton	EKALVSSVRQRMKCS	-	Mus Pep11	EKALVSSVRQRMKCS
16	Singleton	LEQIHVLENSLVL	-	Mus Pep1	LEQIHVLENSLVL
17	Singleton	DDVVASEALNSVWSGF	-	Mus Pep15	DDVVASEALNSVWSGF
18	Singleton	SRPFIFQEVIDLGGE	-	Mus Pep12	SRPFIFQEVIDLGGE
19	Singleton	DKETLSLEELKALL	-	Mus Pep20	DKETLSLEELKALL
20	Singleton	IGGPDDGVITPWQSSF	-	Mus Pep19	IGGPDDGVITPWQSSF
21	Singleton	DIKEKFAKLCEAHGITRDN	-	Rat Pep2	DIKEKFAKLCEAHGITRDN
22	Singleton	RELYLVAYKTPEDGEYFVEY	-	Rat Pep15	RELYLVAYKTPEDGEYFVEY
23	Singleton	ILGKLVKDYHLQFHR	-	Mus Pep18	ILGKLVKDYHLQFHR
24	Singleton	TIFISLFLSVCYSA	-	Mus Pep23	TIFISLFLSVCYSA
25	Singleton	EELRRLAPITSDPTE	-	Mus Pep22	EELRRLAPITSDPTE
26	Singleton	NLDVAKLNGDWFSIVVASNK	-	Rat Pep13	NLDVAKLNGDWFSIVVASNK
27	Singleton	LCEAHGITRDNIDLTKTDR	-	Rat Pep11	LCEAHGITRDNIDLTKTDR
28	Singleton	RIKENGECRELYLVAYKTPE	-	Rat Pep16	RIKENGECRELYLVAYKTPE
29	Singleton	ASNKREKIEENGSMRVFMQH	-	Rat Pep1	ASNKREKIEENGSMRVFMQH
30	Singleton	EEASSTRGNLDVAKLNGDWF	-	Rat Pep3	EEASSTRGNLDVAKLNGDWF

Table 3: Statistics of peptides from MHCLE dataset and their clustering at 70% sequence identity threshold

Features	MHCLE Class I	MHCLE Class II
# Peptides	105642	33757
# Peptides Clustered	57455	28523
# Singletons	48187	5234
Singletons (% Peptides)	46%	16%
# Clusters	11932	4683
Average size of Cluster	4.82	6.09
# Cliques	36732	25941
Average size of Cliques	2.78 ^a	57.36 ^a

^a Number of peptides/Cliques (One peptide can be present in several cliques)

denotes the count of a particular feature

Table 4: Distribution of cluster size for MHCLE dataset for class II with different approaches

Cluster size	Number of clusters		
	Raw Data	Length filtered data ^a (8-25 residues)	Length filtered data ^a + Cluster break algorithm
>=1000	1	0	0
100-999	10	11	10
50-99	21	17	26
30-49	55	50	49
10-29	439	452	492
<10	4157	4196	4335
Total	4683	4726	4912

^a Length filtered data: to obtain a final list of peptide datasets where short (<8 amino acids) and long (> 25 residues) peptides have been removed

Table 5: Distribution of peptide length in both the classes of MHCLE, MHC binding and T cell assays datasets.

Length	MHCLE I (%)	MHCLE II (%)	MHC binding I (%)	MHC binding II (%)	CD8 T cell (%)	CD4 T cell (%)
<8	0.63	0.20	0.16	0.60	0.09	0.24
8	7.04	1.10	6.33	0.25	8.24	0.49
9	50.92	1.37	61.51	2.59	44.30	1.13
10	18.38	2.00	24.85	3.54	20.70	2.33
11	11.74	3.35	4.24	2.56	4.73	1.22
12	4.74	5.85	0.44	2.91	0.64	6.57
13	2.72	10.60	0.19	5.50	0.33	2.76
14	1.36	15.08	0.16	3.01	1.28	2.86
15	0.87	16.49	1.01	56.55	14.61	40.69
16	0.38	14.54	0.18	4.07	0.45	6.11
17	0.30	10.07	0.06	3.43	0.34	3.89
18	0.22	6.33	0.06	3.13	1.09	4.52
19	0.17	3.91	0.02	1.56	0.11	2.12
20	0.12	2.68	0.72	7.62	2.54	18.42
21	0.09	1.84	0.05	1.20	0.05	1.64
22	0.07	1.15	0.01	0.30	0.02	0.56
23	0.05	0.94	0.00	0.14	0.08	0.44
24	0.04	0.65	0.01	0.21	0.05	0.46
25	0.02	0.45	0.00	0.30	0.06	1.05
>25	0.14	1.40	0.00	0.53	0.25	2.52
>=8 and <=25	99.22	98.41	99.83	98.87	99.65	97.24

MHC: Major Histocompatibility Complex, MHC I: MHC class I, MHC II: MHC class II, MHCLE: MHC Ligand Elution, CD8: T cell data recognized by MHC I, CD4: T cell data recognized by MHC II.

Table 6: Distribution of cluster size for different datasets

Cluster size	Number of clusters (Length filtered data + Cluster break algorithm)					
	MHCLE II	MHCLE I	MHC binding I	MHC binding II	CD8 T cell	CD4 T cell
>=1000	0	0	0	0	0	0
100-999	10	0	6	9	0	6
50-99	26	8	24	14	12	23
30-49	49	9	60	32	58	77
10-29	492	277	507	202	431	431
<10	4335	15842	5525	1933	4081	3421
Total	4912	16136	6122	2190	4582	3958

MHC: Major Histocompatibility Complex, MHC I: MHC class I, MHC II: MHC class II, MHCLE: MHC Ligand Elution, CD8: T cell data recognized by MHC I, CD4: T cell data recognized by MHC II.

Table 7: Features and applications of different clustering approaches implemented in Cluster2

Features	Clique	Connected clusters	Cluster-break
Clear consensus sequence	Mostly clear, as peptides are fully interconnected	May have many 'X' at different ambiguous positions	Clusters are broken down to get maximum possible clear consensus
Large membership issue	No	Yes	Resolved
Redundancy in clusters	Yes	No	No
Same sequence identity threshold between peptides in clusters	Yes	No	No
Application	Generating mega pools	Cross-reactivity of small datasets	Cross-reactivity of large datasets

Table 8: Feature-based comparison of different clustering algorithms.

Feature	GibbsCluster	PepServe	Hammock	UCLUST	Cluster1.0	Cluster2.0
Input sequences	Amino acid sequence	Amino acids subjected to retrieval	Amino acid sequence	Amino acid sequence	Amino acid sequence	Amino acid sequence
Freely available online tool	Yes	Yes	Yes (Galaxy)	No	Yes	Yes
Graphical visualization	Yes	Yes	No	No	No	Yes
Provides connectivity in a cluster	No	No	No	No	No	Yes
Cluster representative sequence	Motif	No	Main sequence	No	No	Consensus sequence
Large membership issue	NA	NA	NA	Yes	Yes	Resolved
Overhang sequences identity calculation	NA	NA	NA	Yes, but consider only aligned region	No	Yes
Clustering basis	Supervised	Unsupervised	Unsupervised	Unsupervised	Unsupervised	Unsupervised

NA: feature can't be compared.

List of Figures

Fig 1: Plot representing the number of peptides in top10 clusters from MHC class II ligand elution data

Fig 2: Example cluster visualization before (A) and after (B) cluster-breaking algorithms

Fig 3: Plot Representing the top10 clusters in different datasets after applying cluster-break algorithm

Fig 4: Analysis of overlapping clusters in MHC binding, T cell and MHC ligand elution data. A) H-chart for overlapping clusters between MHC class I binding and CD8 T cell assays. B) H-chart for overlapping clusters between MHC class II binding and CD4 T cell assays. C) Pie-chart of overlapping clusters in MHC class I ligand elution data and CD8 T cell assays. D) Pie-chart of overlapping clusters in MHC class II ligand elution data and CD4 T cell assays. E) Pie-chart of overlapping clusters in MHC class I ligand elution and binding assays data. F) Pie-chart of overlapping clusters in MHC class II ligand elution and binding assays data.

Fig 5: Screenshots from online tool. A) Specify Sequence (Step 1), B) Select clustering parameters (Step 2), C) Choose clustering algorithm (Step 3), D) Tabular output (Result page), E) Graphical Output (Result page)

Supplementary information

S1 Fig: Example representation of clustering using graph theory.

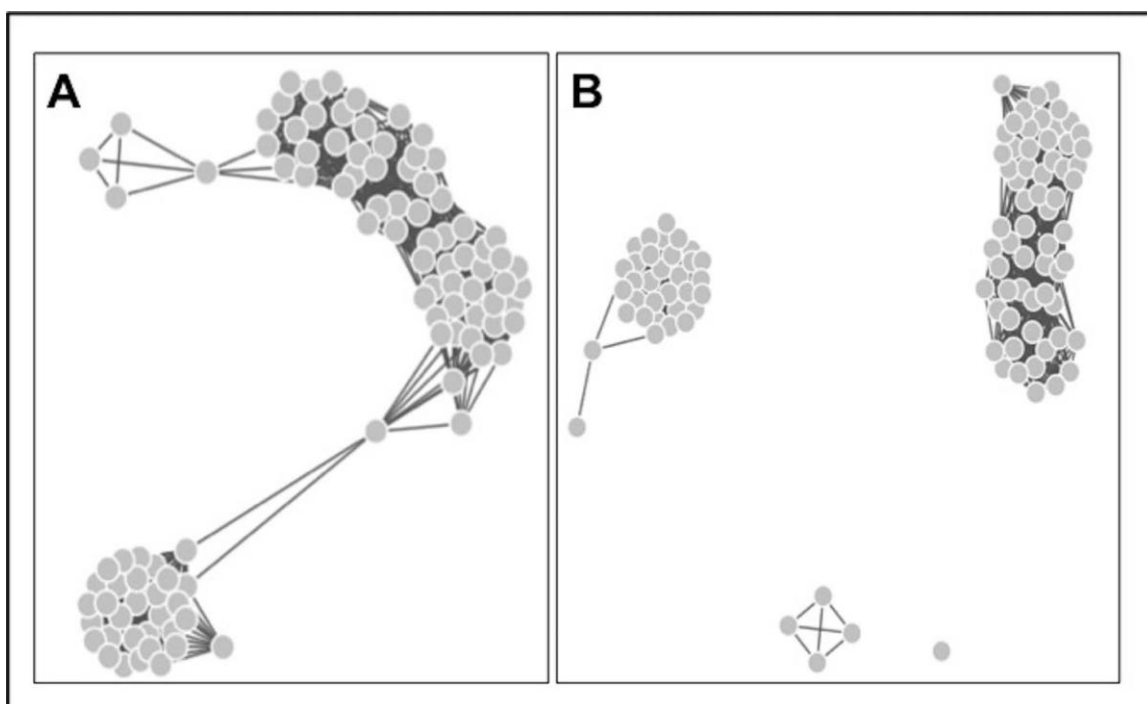
S1 Table: Number of epitopes for MHC Class I and Class II from each dataset.

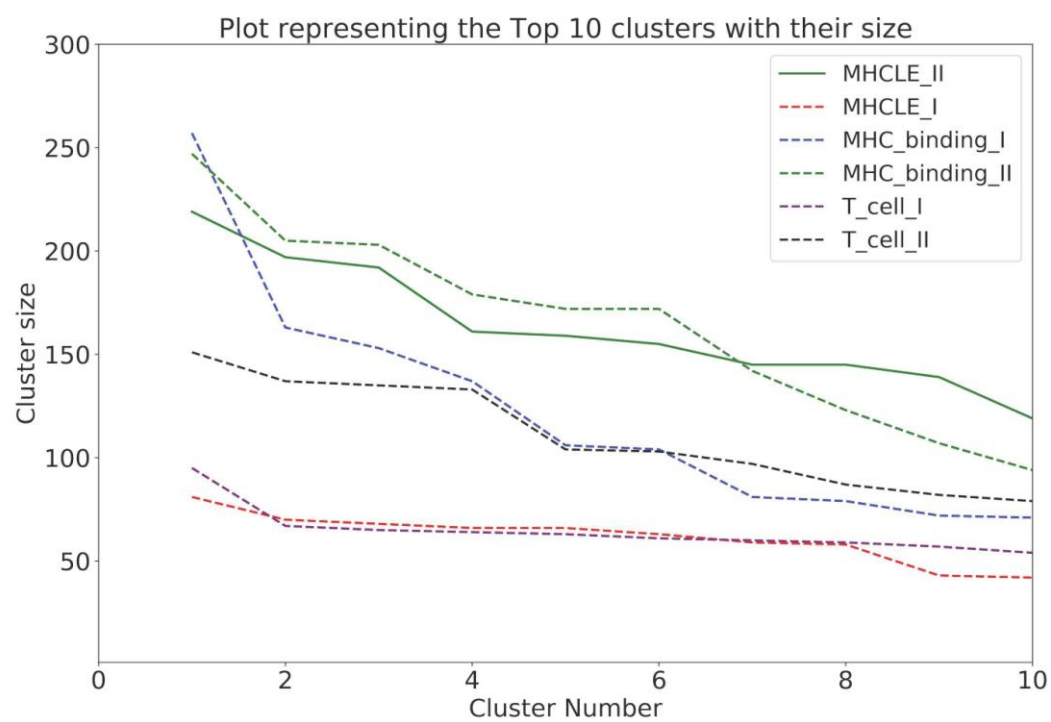
S2 Table: Output of clustering tool for DENV data

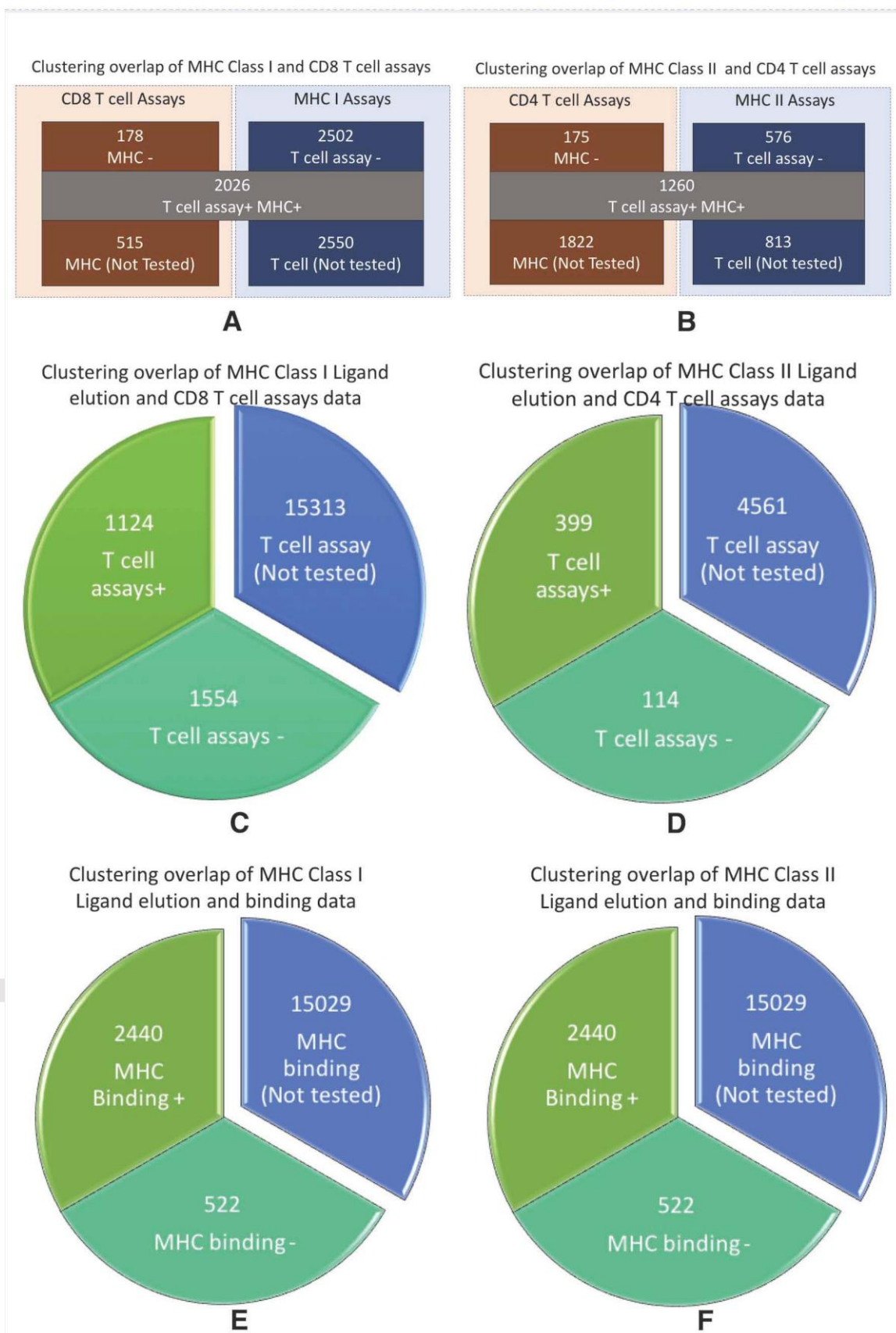
S3 Table: Selected peptides from DENV data in all the cliques and singletons

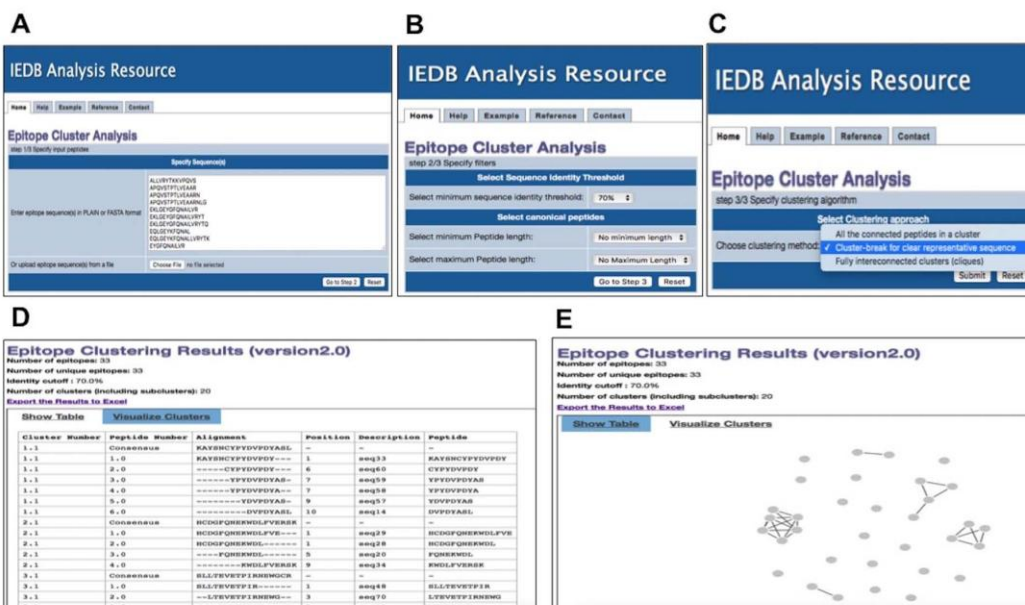
S4 Table: Analysis of overlapping clusters among different dataset











Supplementary information

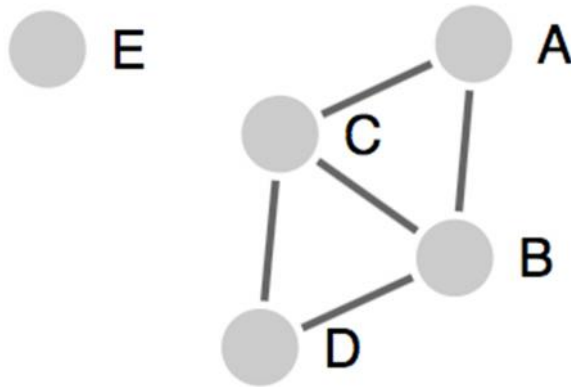
S1 Fig: Example representation of clustering using graph theory.

S1 Table: Number of epitopes for MHC Class I and Class II from each dataset.

S2 Table: Output of clustering tool for DENV data

S3 Table: Selected peptides from DENV data in all the cliques and singletons

S4 Table: Analysis of overlapping clusters among different dataset



S1 Fig: Example representation of clustering using graph theory.

S1 Table: Number of epitopes for MHC Class I and Class II from each dataset.

Dataset	MHC class I	MHC class II
T cell	33,100	37,844
MHCLE	105,642	33,757
MHC binding	44,558	19,611

S2 Table: Output of clustering tool for DENV data

Clique Number	Peptide Number	Alignment	Position	Description	Peptide
1	Consensus	RXLRTLILAPTRVVASEMA	-	-	-
1	1	RGLRTLILAPTRVVA----	1	DENV2_1692_1706 62.7	RGLRTLILAPTRVVA
1	2	RRLRTLILAPTRVVA----	1	DENV3_1692_1706 1036.7	RRLRTLILAPTRVVA
1	3	RKLRTLILAPTRVVA----	1	DENV1_1692_1706 993.3	RKLRTLILAPTRVVA
1	4	-GLRTLILAPTRVVAA---	2	DENV2_1693_1707 184.8	GLRTLILAPTRVVAA
1	5	-KLRTLILAPTRVVAS---	2	DENV1_1693_1707 193.3	KLRTLILAPTRVVAS
1	6	--LRTLILAPTRVVAAE--	3	DENV4_1694_1708 107.7	LRTLILAPTRVVAAE
1	7	--LRTLILAPTRVVASE--	3	DENV1_1694_1708 86.5, DENV1_1694_1708 100.0	LRTLILAPTRVVASE
1	8	---RTLILAPTRVVASEM-	4	DENV1_1695_1709 82.1	RTLILAPTRVVASEM
1	9	----TLILAPTRVVASEMA	5	DENV1_1696_1710 81.3	TLILAPTRVVASEMA
2	Consensus	KRXLRTLILAPTRVVASEM	-	-	-
2	1	KRKLRTLILAPTRVV----	1	DENV1_1691_1705 195.6	KRKLRTLILAPTRVV
2	2	-RGLRTLILAPTRVVA---	2	DENV2_1692_1706 62.7	RGLRTLILAPTRVVA
2	3	-RRLRTLILAPTRVVA---	2	DENV3_1692_1706 1036.7	RRLRTLILAPTRVVA
2	4	-RKLRTLILAPTRVVA---	2	DENV1_1692_1706 993.3	RKLRTLILAPTRVVA
2	5	--GLRTLILAPTRVVAA--	3	DENV2_1693_1707 184.8	GLRTLILAPTRVVAA
2	6	--KLRTLILAPTRVVAS--	3	DENV1_1693_1707 193.3	KLRTLILAPTRVVAS
2	7	---LRTLILAPTRVVASE-	4	DENV1_1694_1708 86.5, DENV1_1694_1708 100.0	LRTLILAPTRVVASE
2	8	---LRTLILAPTRVVAAE-	4	DENV4_1694_1708 107.7	LRTLILAPTRVVAAE
2	9	---RTLILAPTRVVASEM	5	DENV1_1695_1709 82.1	RTLILAPTRVVASEM
3	Consensus	AXIAFLRFLAIPPTAGIL	-	-	-
3	1	ALVAFLRFLTIPPTA----	1	DENV2_50_64 1017.9	ALVAFLRFLTIPPTA
3	2	AFIAFLRFLAIPPTA----	1	DENV1_50_64 458.8	AFIAFLRFLAIPPTA
3	3	-LVAFLRFLTIPPTAG--	2	DENV2_51_65 374.2, DENV2_51_65 88.9	LVAFLRFLTIPPTAG
3	4	-FIAFLRFLAIPPTAG--	2	DENV3_51_65 502.0, DENV1_51_65 265.9, DENV1_51_65 154.0	FIAFLRFLAIPPTAG
3	5	--IAFLRFLAIPPTAGV-	3	DENV3_52_66 853.3	IAFLRFLAIPPTAGV
3	6	--IAFLRFLAIPPTAGI-	3	DENV1_52_66 642.0, DENV1_52_66 123.8, DENV1_52_66 182.9	IAFLRFLAIPPTAGI
3	7	--VAFLRFLTIPPTAGI-	3	DENV2_52_66 579.8, DENV2_52_66 540.6	VAFLRFLTIPPTAGI
3	8	---AFLRFLAIPPTAGIL	4	DENV1_53_67 414.4	AFLRFLAIPPTAGIL
3	9	---AFLRFLTIPPTAGIL	4	DENV2_53_67 44.8	AFLRFLTIPPTAGIL
4	Consensus	XPSFNMLKRXRNRVSTXX	-	-	-
4	1	QPSFNMLKRARNRVS----	1	DENV1_11_25 111.9, DENV1_11_25 64.7	QPSFNMLKRARNRVS
4	2	RPSFNMLKRVNRNRS---	1	DENV1_11_25 313.3	RPSFNMLKRVNRNRS
4	3	-PSFNMLKRARNRVST--	2	DENV1_12_26 242.9	PSFNMLKRARNRVST
4	4	-PSINMLKRVNRNVST--	2	DENV3_12_26 185.3	PSINMLKRVNRNVST
4	5	-TPFNMLKRENRNVST--	2	DENV2_12_26 353.5	TPFNMLKRENRNVST
4	6	---SINMLKRVNRNVSTG-	3	DENV3_13_27 62.5	SINMLKRVNRNVSTG
4	7	---FNMLKRENRNVSTPQ	4	DENV4_14_28 167.9	FNMLKRENRNVSTPQ
4	8	---FNMLKRARNRVSTVS	4	DENV1_14_28 103.3	FNMLKRARNRVSTVS
5	Consensus	RPSFNMLKRXRNRVSTXX	-	-	-
5	1	RPSFNMLKRVNRNRS---	1	DENV1_11_25 313.3	RPSFNMLKRVNRNRS
5	2	-PSFNMLKRARNRVST--	2	DENV1_12_26 242.9	PSFNMLKRARNRVST
5	3	-PSINMLKRVNRNVST--	2	DENV3_12_26 185.3	PSINMLKRVNRNVST
5	4	-TPFNMLKRENRNVST--	2	DENV2_12_26 353.5	TPFNMLKRENRNVST
5	5	---SINMLKRVNRNVSTG-	3	DENV3_13_27 62.5	SINMLKRVNRNVSTG
5	6	---INMLKRVNRNVSTGP	4	DENV3_14_28 31.5	INMLKRVNRNVSTGP
5	7	---FNMLKRENRNVSTPQ	4	DENV4_14_28 167.9	FNMLKRENRNVSTPQ
5	8	---FNMLKRARNRVSTVS	4	DENV1_14_28 103.3	FNMLKRARNRVSTVS
6	Consensus	GVFHTMWHVTRGAVLXX	-	-	-
6	1	GVFHTMWHVTRGSVI--	1	DENV4_1521_1535 339.3	GVFHTMWHVTRGSVI
6	2	GVFHTMWHVTRGAVL--	1	DENV3_1521_1535 524.8, DENV3_1521_1535 221.5	GVFHTMWHVTRGAVL
6	3	GTFTMWHVTRGAVL--	1	DENV2_1521_1535 598.0	GTFTMWHVTRGAVL
6	4	--FHTMWHVTRGAVLMH	3	DENV2_1523_1537 1021.9, DENV2_1523_1537 405.9, DENV2_1523_1537 456.0	FHTMWHVTRGAVLMH
6	5	--FHTMWHVTRGAVLTY	3	DENV3_1523_1537 642.2	FHTMWHVTRGAVLTY

6	6	--FHTMWHVTRGAVLMY	3	DENV1_1523_1537 862.4, DENV1_1523_1537 80.0, DENV1_1523_1537 449.3	FHTMWHVTRGAVLMY
6	7	--FHTMWHVTRGVSICH	3	DENV4_1523_1537 262.8	FHTMWHVTRGVSICH
7	Consensus	FKKEISNMLNIMNRRKRT	-	-	-
7	1	FKKEISNMLNIMNRR---	1	DENV1_85_99 218.1	FKKEISNMLNIMNRR
7	2	FKKEISSMLNIMNRR---	1	DENV1_85_99 407.6	FKKEISSMLNIMNRR
7	3	-KKEISSMLNIMNRRK--	2	DENV1_86_100 87.3	KKEISSMLNIMNRRK
7	4	--KEVSNMLNIMNRRKR-	3	DENV1_87_101 368.8	KEVSNMLNIMNRRKR
7	5	--KEISSMLNIMNRRKR-	3	DENV1_87_101 126.7	KEISSMLNIMNRRKR
7	6	--KEISNMLNIMNRRKR-	3	DENV1_87_101 980.8	KEISNMLNIMNRRKR
7	7	---EISNMLNIMNRRKRT	4	DENV1_88_102 210.6	EISNMLNIMNRRKRT
8	Consensus	FKKEISNMLNIMNRRKRXV	-	-	-
8	1	FKKEISNMLNIMNRR----	1	DENV1_85_99 218.1	FKKEISNMLNIMNRR
8	2	-KKEISSMLNIMNRRK----	2	DENV1_86_100 87.3	KKEISSMLNIMNRRK
8	3	--KEVSNMLNIMNRRKR--	3	DENV1_87_101 368.8	KEVSNMLNIMNRRKR
8	4	--KEISSMLNIMNRRKR--	3	DENV1_87_101 126.7	KEISSMLNIMNRRKR
8	5	--KEISNMLNIMNRRKR--	3	DENV1_87_101 980.8	KEISNMLNIMNRRKR
8	6	---EISNMLNIMNRRKRT-	4	DENV1_88_102 210.6	EISNMLNIMNRRKRT
8	7	----ISNMLNIMNRRKRSV	5	DENV1_89_103 208.2	ISNMLNIMNRRKRSV
9	Consensus	RKNGKKVIQLSRKTFDTEY	-	-	-
9	1	RKNGKKVIQLSRKTF----	1	DENV3_1854_1868 48.9, DENV2_1854_1868 675.3	RKNGKKVIQLSRKTF
9	2	--SGKKVIQLSRKTFDT--	3	DENV4_1856_1870 1065.5	SGKKVIQLSRKTFDT
9	3	--NGKRVIQLSRKTFDT--	3	DENV1_1856_1870 467.1	NGKRVIQLSRKTFDT
9	4	--NGKKVIQLSRKTFDS--	3	DENV2_1856_1870 141.1	NGKKVIQLSRKTFDS
9	5	--NGKKVIQLSRKTFDT--	3	DENV3_1856_1870 442.4	NGKKVIQLSRKTFDT
9	6	---KKVIQLSRKTFDSEY	5	DENV2_1858_1872 125.2	KKVIQLSRKTFDSEY
10	Consensus	MALVAFLRFLTIPPTAGIL	-	-	-
10	1	MALVAFLRFLTIPPT-----	1	DENV2_49_63 179.2, DENV2_49_63 50.0	MALVAFLRFLTIPPT
10	2	-ALVAFLRFLTIPPTA---	2	DENV2_50_64 1017.9	ALVAFLRFLTIPPTA
10	3	-AFIAFLRFLAIPPTA---	2	DENV1_50_64 458.8	AFIAFLRFLAIPPTA
10	4	--LVAFLRFLTIPPTAG--	3	DENV2_51_65 374.2, DENV2_51_65 88.9	LVAFLRFLTIPPTAG
10	5	---VAFLRFLTIPPTAGI-	4	DENV2_52_66 579.8, DENV2_52_66 540.6	VAFLRFLTIPPTAGI
10	6	----AFLRFLTIPPTAGIL	5	DENV2_53_67 44.8	AFLRFLTIPPTAGIL
11	Consensus	GLRTLILAPTRVVAXEMXE	-	-	-
11	1	GLRTLILAPTRVVAA-----	1	DENV2_1693_1707 184.8	GLRTLILAPTRVVAA
11	2	-LRTLILAPTRVVASE---	2	DENV1_1694_1708 86.5, DENV1_1694_1708 100.0	LRTLILAPTRVVASE
11	3	-LRTLILAPTRVVAAE----	2	DENV4_1694_1708 107.7	LRTLILAPTRVVAAE
11	4	--RTLILAPTRVVASEM--	3	DENV1_1695_1709 82.1	RTLILAPTRVVASEM
11	5	---TLILAPTRVVASEMA-	4	DENV1_1696_1710 81.3	TLILAPTRVVASEMA
11	6	----LILAPTRVVAEMEE	5	DENV3_1697_1711 488.7	LILAPTRVVAEMEE
12	Consensus	XPSFNMLKRVNRNVSTG	-	-	-
12	1	QPSFNMLKRVNRNVS--	1	DENV1_11_25 111.9, DENV1_11_25 64.7	QPSFNMLKRVNRNVS
12	2	KPSINMLKRVNRNVS--	1	DENV3_11_25 105.9	KPSINMLKRVNRNVS
12	3	RPSFNMLKRVNRNVS--	1	DENV1_11_25 313.3	RPSFNMLKRVNRNVS
12	4	-PSFNMLKRVNRNVST-	2	DENV1_12_26 242.9	PSFNMLKRVNRNVST
12	5	-PSINMLKRVNRNVST-	2	DENV3_12_26 185.3	PSINMLKRVNRNVST
12	6	--SINMLKRVNRNVSTG	3	DENV3_13_27 62.5	SINMLKRVNRNVSTG
13	Consensus	XPSINMLKRVNRNVSTGP	-	-	-
13	1	KPSINMLKRVNRNVS----	1	DENV3_11_25 105.9	KPSINMLKRVNRNVS
13	2	RPSFNMLKRVNRNVS---	1	DENV1_11_25 313.3	RPSFNMLKRVNRNVS
13	3	-PSFNMLKRVNRNVST--	2	DENV1_12_26 242.9	PSFNMLKRVNRNVST
13	4	-PSINMLKRVNRNVST--	2	DENV3_12_26 185.3	PSINMLKRVNRNVST
13	5	---SINMLKRVNRNVSTG-	3	DENV3_13_27 62.5	SINMLKRVNRNVSTG
13	6	----INMLKRVNRNVSTGP	4	DENV3_14_28 31.5	INMLKRVNRNVSTGP
14	Consensus	KXXIFILLMLVTPSMAMR	-	-	-
14	1	KVVIFILLMLVTPSM---	1	DENV3_266_280 222.5	KVVIFILLMLVTPSM
14	2	-GIIFILLMLVTPSMA--	2	DENV1_267_281 58.0	GIIFILLMLVTPSMA
14	3	--VIFILLMLVTPSMTM-	3	DENV3_268_282 101.7	VIFILLMLVTPSMTM
14	4	---IVFILLMLVTPSMAM-	3	DENV1_268_282 478.5	IVFILLMLVTPSMAM
14	5	----IFILLMLVTPSMAMR	4	DENV1_269_283 133.3	IFILLMLVTPSMAMR
15	Consensus	NRQDLLVTFKTAHAKK	-	-	-
15	1	NRQDLLVTFKTAHAK-	1	DENV1_513_527 382.9	NRQDLLVTFKTAHAK

15	2	NRKELLVTFKNAHAK-	1	DENV3_513_527 361.1	NRKELLVTFKNAHAK
15	3	-RKELLVTFKNAHAKK	2	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
15	4	-RKDLLVTFKTAHAKK	2	DENV1_514_528 104.0	RKDLLVTFKTAHAKK
15	5	-REDLLVTFKTAHAKK	2	DENV1_514_528 323.8	REDLLVTFKTAHAKK
16	Consensus	NRXDLLVTFKTAHAKKQ	-	-	-
16	1	NRQDLLVTFKTAHAK--	1	DENV1_513_527 382.9	NRQDLLVTFKTAHAK
16	2	-RKELLVTFKNAHAKK-	2	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
16	3	-RKDLLVTFKTAHAKK-	2	DENV1_514_528 104.0	RKDLLVTFKTAHAKK
16	4	-REDLLVTFKTAHAKK-	2	DENV1_514_528 323.8	REDLLVTFKTAHAKK
16	5	--QDLLVTFKTAHAKKQ	3	DENV1_515_529 58.0, DENV1_515_529 67.6	QDLLVTFKTAHAKKQ
17	Consensus	NRKELLVTFKNAHAKKQE	-	-	-
17	1	NRKELLVTFKNAHAK---	1	DENV3_513_527 361.1	NRKELLVTFKNAHAK
17	2	-REDLLVTFKTAHAKK--	2	DENV1_514_528 323.8	REDLLVTFKTAHAKK
17	3	-RKELLVTFKNAHAKK--	2	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
17	4	-RKDLLVTFKTAHAKK--	2	DENV1_514_528 104.0	RKDLLVTFKTAHAKK
17	5	---ELLVTFKNAHAKKQE	4	DENV3_516_530 656.0, DENV3_516_530 166.1, DENV3_516_530 56.7	ELLVTFKNAHAKKQE
18	Consensus	RXDLLVTFKTAHAKKQE	-	-	-
18	1	RKELLVTFKNAHAKK--	1	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
18	2	REDLLVTFKTAHAKK--	1	DENV1_514_528 323.8	REDLLVTFKTAHAKK
18	3	RKDLLVTFKTAHAKK--	1	DENV1_514_528 104.0	RKDLLVTFKTAHAKK
18	4	-QDLLVTFKTAHAKKQ-	2	DENV1_515_529 58.0, DENV1_515_529 67.6	QDLLVTFKTAHAKKQ
18	5	--ELLVTFKNAHAKKQE	3	DENV3_516_530 656.0, DENV3_516_530 166.1, DENV3_516_530 56.7	ELLVTFKNAHAKKQE
19	Consensus	KEIGRMLNILNRRRRRTAG	-	-	-
19	1	KEIGRMLNILNRRRR---	1	DENV2_87_101 651.3, DENV2_87_101 68.0	KEIGRMLNILNRRRR
19	2	-EIGRMLNILNKRRT--	2	DENV2_88_102 198.7	EIGRMLNILNKRRT
19	3	---IGRMLNILSRRRRTA-	3	DENV2_89_103 199.8	IGRMLNILSRRRRTA
19	4	---IGRMLNILNRRRRTA-	3	DENV2_89_103 260.8, DENV2_89_103 309.4	IGRMLNILNRRRRTA
19	5	---GRMLNIMNRRRRRTAG	4	DENV2_90_104 474.3	GRMLNIMNRRRRRTAG
20	Consensus	ILAKAIFKLTYNKVVVRV	-	-	-
20	1	ILAKAIFKLTYNQKV---	1	DENV4_3058_3072 165.7	ILAKAIFKLTYNQKV
20	2	-LAKAIFKLTYNQKV--	2	DENV4_3059_3073 870.7	LAKAIFKLTYNQKV
20	3	---ATSIKLTYNQKVVVR-	3	DENV1_3060_3074 866.0	ATSIKLTYNQKVVVR
20	4	---EAIKLTYNKVVVRV	4	DENV2_3061_3075 655.0	EAIKLTYNKVVVRV
20	5	---KAIFKLTYNKVVVKV	4	DENV4_3061_3075 655.0	KAIFKLTYNKVVVKV
21	Consensus	LAXAIFKLTYNKVVVRVQ	-	-	-
21	1	LAKAIFKLTYNQKV---	1	DENV4_3059_3073 870.7	LAKAIFKLTYNQKV
21	2	-ATSIKLTYNQKVVVR--	2	DENV1_3060_3074 866.0	ATSIKLTYNQKVVVR
21	3	---KAIFKLTYNKVVVKV-	3	DENV4_3061_3075 655.0	KAIFKLTYNKVVVKV
21	4	---EAIKLTYNKVVVRV-	3	DENV2_3061_3075 655.0	EAIKLTYNKVVVRV
21	5	---SIFKLTYNKVVVRVQ	4	DENV1_3062_3076 687.5	SIFKLTYNKVVVRVQ
22	Consensus	FMALVAFLRFLTIPPTAGI	-	-	-
22	1	FMALVAFLRFLTIPP----	1	DENV2_48_62 151.7	FMALVAFLRFLTIPP
22	2	-MALVAFLRFLTIPPT----	2	DENV2_49_63 179.2, DENV2_49_63 50.0	MALVAFLRFLTIPPT
22	3	---ALVAFLRFLTIPPTA--	3	DENV2_50_64 1017.9	ALVAFLRFLTIPPTA
22	4	---LVAFLRFLTIPPTAG-	4	DENV2_51_65 374.2, DENV2_51_65 88.9	LVAFLRFLTIPPTAG
22	5	----VAFLRFLTIPPTAGI	5	DENV2_52_66 579.8, DENV2_52_66 540.6	VAFLRFLTIPPTAGI
23	Consensus	WSLMYFHRDLRLAANA	-	-	-
23	1	WTLMYFHRDLRLAA--	1	DENV2_3258_3272 252.7, DENV2_3258_3272 54.0	WTLMYFHRDLRLAA
23	2	WSLMYFHRDLRLAA--	1	DENV2_3258_3272 125.2, DENV2_3258_3272 129.4	WSLMYFHRDLRLAA
23	3	-SLMYFHRDLRLASN-	2	DENV3_3259_3273 349.8	SLMYFHRDLRLASN
23	4	---LMYFHRDLRLAANA	3	DENV2_3260_3274 446.5	LMYFHRDLRLAANA
23	5	---LMYFHRDLRLASNA	3	DENV3_3260_3274 417.7, DENV3_3260_3274 64.7	LMYFHRDLRLASNA
24	Consensus	KNGAIKVLRGFKKEISN	-	-	-
24	1	KNGAIKVLRGFKKEI--	1	DENV1_75_89 622.7	KNGAIKVLRGFKKEI

24	2	KSGAIKVLKGFKKEI--	1	DENV3_75_89 817.1	KSGAIKVLKGFKKEI
24	3	-NGAIKVLRGFKKEIS-	2	DENV1_76_90 752.9, DENV1_76_90 69.3	NGAIKVLRGFKKEIS
24	4	--GAIKVLKGFKKEISN	3	DENV3_77_91 545.0	GAIKVLKGFKKEISN
24	5	--GAIKVLRGFKKEISN	3	DENV1_77_91 270.8, DENV1_77_91 120.7	GAIKVLRGFKKEISN
25	Consensus	FKKNGAIKVLRGFKKEISN	-	-	-
25	1	FKKNGAIKVLRGFKK----	1	DENV1_73_87 117.0, DENV1_73_87 1471.8, DENV1_73_87 838.7	FKKNGAIKVLRGFKK
25	2	--KNGAIKVLRGFKKEI--	3	DENV1_75_89 622.7	KNGAIKVLRGFKKEI
25	3	--KSGAIKVLKGFKKEI--	3	DENV3_75_89 817.1	KSGAIKVLKGFKKEI
25	4	---NGAIKVLRGFKKEIS-	4	DENV1_76_90 752.9, DENV1_76_90 69.3	NGAIKVLRGFKKEIS
25	5	----GAIKVLRGFKKEISN	5	DENV1_77_91 270.8, DENV1_77_91 120.7	GAIKVLRGFKKEISN
26	Consensus	VMAFIAFLRFLAIPPTAGX	-	-	-
26	1	VMAFIAFLRFLAIPP----	1	DENV1_48_62 229.2, DENV1_48_62 153.3, DENV3_48_62 56.0	VMAFIAFLRFLAIPP
26	2	--AFIAFLRFLAIPPTA--	3	DENV1_50_64 458.8	AFIAFLRFLAIPPTA
26	3	---FIAFLRFLAIPPTAG-	4	DENV3_51_65 502.0, DENV1_51_65 265.9, DENV1_51_65 154.0	FIAFLRFLAIPPTAG
26	4	----IAFLRFLAIPPTAGI	5	DENV1_52_66 642.0, DENV1_52_66 123.8, DENV1_52_66 182.9	IAFLRFLAIPPTAGI
26	5	----IAFLRFLAIPPTAGV	5	DENV3_52_66 853.3	IAFLRFLAIPPTAGV
27	Consensus	VLGRFRKEIGRMLNILNXR	-	-	-
27	1	VLGRFRKEIGRMLNI----	1	DENV2_81_95 328.6	VLGRFRKEIGRMLNI
27	2	-LRGFRKEIGRMLNIL----	2	DENV2_82_96 315.2	LRGFRKEIGRMLNIL
27	3	--IGFRKEIGRMLNILN--	3	DENV4_83_97 69.4, DENV4_83_97 441.9	IGFRKEIGRMLNILN
27	4	----FRKEIGRMLNILNGR	5	DENV4_85_99 260.0, DENV4_85_99 561.9	FRKEIGRMLNILNGR
27	5	----FRKEIGRMLNILNRR	5	DENV2_85_99 323.5, DENV2_85_99 238.1, DENV2_85_99 616.2	FRKEIGRMLNILNRR
28	Consensus	VLRGFXKEIXXMLNIX	-	-	-
28	1	VLGRFRKEIGRMLNI-	1	DENV2_81_95 328.6	VLGRFRKEIGRMLNI
28	2	VLRGFKKEISNMLNI-	1	DENV1_81_95 182.1, DENV1_81_95 456.7	VLRGFKKEISNMLNI
28	3	-LRGFRKEIGRMLNIL	2	DENV2_82_96 315.2	LRGFRKEIGRMLNIL
28	4	-LRGFKKEISSMLNIM	2	DENV1_82_96 377.1	LRGFKKEISSMLNIM
29	Consensus	XLAXXIFKLTQYQNKVVR	-	-	-
29	1	ILAKAIFKLTQYQNKV--	1	DENV4_3058_3072 165.7	ILAKAIFKLTQYQNKV
29	2	LLATSIFKLTQYQNKV--	1	DENV1_3058_3072 385.9	LLATSIFKLTQYQNKV
29	3	-LAKAIFKLTQYQNKVV-	2	DENV4_3059_3073 870.7	LAKAIFKLTQYQNKVV
29	4	--ATSIFKLTQYQNKVVR	3	DENV1_3060_3074 866.0	ATSIFKLTQYQNKVVR
30	Consensus	LLATSIFKLTQYQNKVVRVQ	-	-	-
30	1	LLATSIFKLTQYQNKV----	1	DENV1_3058_3072 385.9	LLATSIFKLTQYQNKV
30	2	-LAKAIFKLTQYQNKVV----	2	DENV4_3059_3073 870.7	LAKAIFKLTQYQNKVV
30	3	--ATSIFKLTQYQNKVVR--	3	DENV1_3060_3074 866.0	ATSIFKLTQYQNKVVR
30	4	----SIFKLTQYQNKVVRVQ	5	DENV1_3062_3076 687.5	SIFKLTQYQNKVVRVQ
31	Consensus	KLFMALVAFLRFLTIPPTA	-	-	-
31	1	KLFMALVAFLRFLTII----	1	DENV2_46_60 192.3, DENV2_46_60 151.1	KLFMALVAFLRFLTI
31	2	--FMALVAFLRFLTIPP--	3	DENV2_48_62 151.7	FMALVAFLRFLTIPP
31	3	---MALVAFLRFLTIPPT-	4	DENV2_49_63 179.2, DENV2_49_63 50.0	MALVAFLRFLTIPPT
31	4	----ALVAFLRFLTIPPTA	5	DENV2_50_64 1017.9	ALVAFLRFLTIPPTA
32	Consensus	LRGFKKEISSMLNIMNRRK	-	-	-
32	1	LRGFKKEISSMLNIM----	1	DENV1_82_96 377.1	LRGFKKEISSMLNIM
32	2	---FKKEISSMLNIMNRR-	4	DENV1_85_99 407.6	FKKEISSMLNIMNRR
32	3	---FKKEISSMLNIMNRR-	4	DENV1_85_99 218.1	FKKEISSMLNIMNRR
32	4	----KKEISSMLNIMNRRK	5	DENV1_86_100 87.3	KKEISSMLNIMNRRK
33	Consensus	XGKVVGLYGNGVVTX	-	-	-
33	1	KGKVVGLYGNGVVTR	1	DENV2_1620_1634 1200.0	KGKVVGLYGNGVVTR
33	2	KGKVIGLYGNGVVTK	1	DENV4_1620_1634 1033.3	KGKVIGLYGNGVVTK
33	3	EGKVVGLYGNGVVTK	1	DENV3_1620_1634 1544.4	EGKVVGLYGNGVVTK
33	4	EGKIVGLYGNGVVT	1	DENV1_1620_1634 1645.0	EGKIVGLYGNGVVT
34	Consensus	TGNIVSSVNMVSRLLNRF	-	-	-
34	1	TGNIVSSVNMVSRLL----	1	DENV3_2720_2734 32.5	TGNIVSSVNMVSRLL
34	2	-GNIVSSVNMVSRLLL----	2	DENV3_2721_2735 29.3	GNIVSSVNMVSRLLL
34	3	---IVASVNMVSRLLLNRF	4	DENV3_2723_2737 261.2	IVASVNMVSRLLLNRF
34	4	----VSSVNMVSRLLLNRF	5	DENV3_2724_2738 38.7	VSSVNMVSRLLLNRF

35	Consensus	TGNIVSSVNMXSRL	-	-	-
35	1	TGNIVSSVNMVSRL--	1	DENV3_2720_2734 32.5	TGNIVSSVNMVSRL
35	2	TGNIVSAVNMTSRML--	1	DENV1_2720_2734 22.3	TGNIVSAVNMTSRML
35	3	SGNIVSSVNMISRL--	1	DENV2_2720_2734 196.5	SGNIVSSVNMISRL
35	4	--GNIVSSVNMVSRL	2	DENV3_2721_2735 29.3	GNIVSSVNMVSRL
36	Consensus	GTGNIVSVNMXSRL	-	-	-
36	1	GTGNIVSAVNMTSRM--	1	DENV1_2719_2733 73.7	GTGNIVSAVNMTSRM
36	2	--SGNIVSSVNMISRL	2	DENV2_2720_2734 196.5	SGNIVSSVNMISRL
36	3	--TGNIVSSVNMVSRL	2	DENV3_2720_2734 32.5	TGNIVSSVNMVSRL
36	4	--TGNIVSAVNMTSRML	2	DENV1_2720_2734 22.3	TGNIVSAVNMTSRML
37	Consensus	LDDGIYRILQRGLGRSQV	-	-	-
37	1	LDDGIYRILQRGLLG----	1	DENV1_1495_1509 80.9	LDDGIYRILQRGLLG
37	2	---DGIYRIMQRGLGRS---	3	DENV1_1497_1511 47.4	DGIYRIMQRGLGRS
37	3	---GIYRILQRGLGRSQ-	4	DENV1_1498_1512 151.1	GIYRILQRGLGRSQ
37	4	----IYRILQRGLGRSQV	5	DENV1_1499_1513 46.1	IYRILQRGLGRSQV
38	Consensus	GVTYLALLAAFKVRPTFAA	-	-	-
38	1	GVTYLALLAAFKVRP----	1	DENV2_1199_1213 167.6	GVTYLALLAAFKVRP
38	2	--VTYLALLAAFKVRPT----	2	DENV2_1200_1214 377.7	VTYLALLAAFKVRPT
38	3	---YLALLAAFKVRPTFA-	4	DENV2_1202_1216 268.0	YLALLAAFKVRPTFA
38	4	----LALLAAFKVRPTFAA	5	DENV2_1203_1217 170.7	LALLAAFKVRPTFAA
39	Consensus	FRKEIGRMLNILNRRRTA	-	-	-
39	1	FRKEIGRMLNILNR----	1	DENV2_85_99 323.5, DENV2_85_99 238.1, DENV2_85_99 616.2	FRKEIGRMLNILNRR
39	2	--KEIGRMLNILNRRR--	3	DENV2_87_101 651.3, DENV2_87_101 68.0	KEIGRMLNILNRRR
39	3	---EIGRMLNILNKRRT-	4	DENV2_88_102 198.7	EIGRMLNILNKRRT
39	4	----IGRMLNILNRRRTA	5	DENV2_89_103 260.8, DENV2_89_103 309.4	IGRMLNILNRRRTA
40	Consensus	RKEXLVTFKNXHAQQE	-	-	-
40	1	RKELLVTFKNAHAKK--	1	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
40	2	--KTLVTFKNPHAKKQ-	2	DENV2_515_529 158.7	KTLVTFKNPHAKKQ
40	3	--KETLVTFKNPHAKKQ-	2	DENV2_515_529 81.0	KETLVTFKNPHAKKQ
40	4	---ELLVTFKNAHAKKQE	3	DENV3_516_530 656.0, DENV3_516_530 166.1, DENV3_516_530 56.7	ELLVTFKNAHAKKQE
41	Consensus	KETLVTFKNPHAKKQXV	-	-	-
41	1	KKTLVTFKNPHAKKQ--	1	DENV2_515_529 158.7	KKTLVTFKNPHAKKQ
41	2	KETLVTFKNPHAKKQ--	1	DENV2_515_529 81.0	KETLVTFKNPHAKKQ
41	3	--ELLVTFKNAHAKKQE-	2	DENV3_516_530 656.0, DENV3_516_530 166.1, DENV3_516_530 56.7	ELLVTFKNAHAKKQE
41	4	---TLVTFKNPHAKKQDV	3	DENV2_517_531 26.0	TLVTFKNPHAKKQDV
42	Consensus	NRKELLVTFKNAHAKKQE	-	-	-
42	1	NRKELLVTFKNAHAK----	1	DENV3_513_527 361.1	NRKELLVTFKNAHAK
42	2	--RKELLVTFKNAHAKK--	2	DENV3_514_528 475.0, DENV3_514_528 202.0	RKELLVTFKNAHAKK
42	3	---KETLVTFKNPHAKKQ-	3	DENV2_515_529 81.0	KETLVTFKNPHAKKQ
42	4	---ELLVTFKNAHAKKQE	4	DENV3_516_530 656.0, DENV3_516_530 166.1, DENV3_516_530 56.7	ELLVTFKNAHAKKQE
43	Consensus	VTYLALLAAFKVRPTFAAG	-	-	-
43	1	VTYLALLAAFKVRPT----	1	DENV2_1200_1214 377.7	VTYLALLAAFKVRPT
43	2	---YLALLAAFKVRPTFA--	3	DENV2_1202_1216 268.0	YLALLAAFKVRPTFA
43	3	----LALLAAFKVRPTFAA-	4	DENV2_1203_1217 170.7	LALLAAFKVRPTFAA
43	4	----ALLAAFKVRPTFAAG	5	DENV2_1204_1218 25.9	ALLAAFKVRPTFAAG
44	Consensus	FRKEIGRMLNILNXRRRT	-	-	-
44	1	FRKEIGRMLNILNGR---	1	DENV4_85_99 260.0, DENV4_85_99 561.9	FRKEIGRMLNILNGR
44	2	FRKEIGRMLNILNRR---	1	DENV2_85_99 323.5, DENV2_85_99 238.1, DENV2_85_99 616.2	FRKEIGRMLNILNRR
44	3	--KEIGRMLNILNRRR--	3	DENV2_87_101 651.3, DENV2_87_101 68.0	KEIGRMLNILNRRR
44	4	---EIGRMLNILNKRRT	4	DENV2_88_102 198.7	EIGRMLNILNKRRT
45	Consensus	IGFRKEIGRMLNILNRRRR	-	-	-
45	1	IGFRKEIGRMLNILN----	1	DENV4_83_97 69.4, DENV4_83_97 441.9	IGFRKEIGRMLNILN
45	2	--FRKEIGRMLNILNGR--	3	DENV4_85_99 260.0, DENV4_85_99 561.9	FRKEIGRMLNILNGR
45	3	--FRKEIGRMLNILNRR--	3	DENV2_85_99 323.5, DENV2_85_99 238.1, DENV2_85_99 616.2	FRKEIGRMLNILNRR

45	4	---KEIGRMLNILNRRRR	5	DENV2_87_101 651.3, DENV2_87_101 68.0	KEIGRMLNILNRRRR
46	Consensus	VLRGFKKEISNMLNIX	-	-	-
46	1	VLRGFKKEISNMLNI-	1	DENV1_81_95 182.1, DENV1_81_95 456.7	VLRGFKKEISNMLNI
46	2	-LKGFKKEISNMLSII	2	DENV3_82_96 135.0, DENV3_82_96 63.0, DENV3_82_96 102.9	LKGFKKEISNMLSII
46	3	-LRGFKKEISSMLNIM	2	DENV1_82_96 377.1	LRGFKKEISSMLNIM
47	Consensus	VLRGFKKEISNMLNIMNRR	-	-	-
47	1	VLRGFKKEISNMLNI----	1	DENV1_81_95 182.1, DENV1_81_95 456.7	VLRGFKKEISNMLNI
47	2	-LRGFKKEISSMLNIM---	2	DENV1_82_96 377.1	LRGFKKEISSMLNIM
47	3	----FKKEISNMLNIMNRR	5	DENV1_85_99 218.1	FKKEISNMLNIMNRR
48	Consensus	AQMWQLMYFHRRDLRLAA	-	-	-
48	1	AQMWQLMYFHRRDLR---	1	DENV1_3255_3269 534.0	AQMWQLMYFHRRDLR
48	2	---WTLMYFHRRDLRLAA	4	DENV2_3258_3272 252.7, DENV2_3258_3272 54.0	WTLMYFHRRDLRLAA
48	3	---WSLMYFHRRDLRLAA	4	DENV2_3258_3272 125.2, DENV2_3258_3272 129.4	WSLMYFHRRDLRLAA
49	Consensus	YAQMWSLMYFHRRDLRLAA	-	-	-
49	1	YAQMWSLMYFHRRDL----	1	DENV4_3254_3268 345.3	YAQMWSLMYFHRRDL
49	2	-AQMWQLMYFHRRDLR---	2	DENV1_3255_3269 534.0	AQMWQLMYFHRRDLR
49	3	---WSLMYFHRRDLRLAA	5	DENV2_3258_3272 125.2, DENV2_3258_3272 129.4	WSLMYFHRRDLRLAA
50	Consensus	KEISXMLNIMNRRKR	-	-	-
50	1	KEISSMLNIMNRRKR	1	DENV1_87_101 126.7	KEISSMLNIMNRRKR
50	2	KEISNMLNIMNRRKR	1	DENV1_87_101 980.8	KEISNMLNIMNRRKR
50	3	KEIGRMLNILNRRRR	1	DENV2_87_101 651.3, DENV2_87_101 68.0	KEIGRMLNILNRRRR
51	Consensus	XGKTVWFVPSIKAGN	-	-	-
51	1	QGKTVWFVPSIKAGN	1	DENV4_1833_1847 197.8	QGKTVWFVPSIKAGN
51	2	AGKTVWFVPSIKAGN	1	DENV3_1833_1847 355.6	AGKTVWFVPSIKAGN
51	3	KGKTVWFVPSIKAGN	1	DENV2_1833_1847 377.0, DENV2_1833_1847 181.3	KGKTVWFVPSIKAGN
52	Consensus	YVLGVFLRKLTSETAL	-	-	-
52	1	YVLGVFLRKLTSET--	1	DENV4_1215_1229 70.8	YVLGVFLRKLTSET
52	2	--LGVFLRKLTSETAL	3	DENV4_1217_1231 57.8	LGVFLRKLTSETAL
52	3	--LGIFLRKLTSETAL	3	DENV4_1217_1231 115.6	LGIFLRKLTSETAL
53	Consensus	IVXSVNMVSRLLLNRFMT	-	-	-
53	1	IVASVNMVSRLLLN---	1	DENV3_2723_2737 261.2	IVASVNMVSRLLLN
53	2	-VSSVNMVSRLLLNRF--	2	DENV3_2724_2738 38.7	VSSVNMVSRLLLNRF
53	3	---SVNMVSRLLLNRFMT	4	DENV3_2726_2740 297.4, DENV3_2726_2740 66.0	SVNMVSRLLLNRFMT
54	Consensus	GMGVTYLALLAAFKVRPT	-	-	-
54	1	GMGVTYLALLAAYKV---	1	DENV2_1197_1211 278.7	GMGVTYLALLAAYKV
54	2	--GVTYLALLAAFKVRP-	3	DENV2_1199_1213 167.6	GVTYLALLAAFKVRP
54	3	---VTYLALLAAFKVRPT	4	DENV2_1200_1214 377.7	VTYLALLAAFKVRPT
55	Consensus	KEISNMLSIIINKRKTSL	-	-	-
55	1	KEISNMLSIIINKRKK---	1	DENV3_87_101 700.0, DENV3_87_101 180.7	KEISNMLSIIINKRKK
55	2	---ISNMLSIIINKRKTS-	3	DENV3_89_103 327.1	ISNMLSIIINKRKTS
55	3	---SNMLSIIINKRKTSL	4	DENV3_90_104 153.0	SNMLSIIINKRKTSL
56	Consensus	GEAAAI FMTATPPGTADA	-	-	-
56	1	GEAAAI FMTATPPGS---	1	DENV1_1785_1799 108.6	GEAAAI FMTATPPGS
56	2	-EAAAI FMTATPPGTA--	2	DENV3_1786_1800 193.3	EAAAI FMTATPPGTA
56	3	---AAI FMTATPPGTADA	4	DENV3_1788_1802 248.6	AAI FMTATPPGTADA
57	Consensus	FKKEISXMLNIMNRR	-	-	-
57	1	FKKEISSMLNIMNRR	1	DENV1_85_99 407.6	FKKEISSMLNIMNRR
57	2	FRKEIGRMLNILNRR	1	DENV2_85_99 323.5, DENV2_85_99 238.1, DENV2_85_99 616.2	FRKEIGRMLNILNRR
57	3	FKKEISNMLNIMNRR	1	DENV1_85_99 218.1	FKKEISNMLNIMNRR
58	Consensus	ALRGLPIRYQTPAIRAE	-	-	-
58	1	ALRGLPIRYQTPAVK--	1	DENV4_1712_1726 412.5	ALRGLPIRYQTPAVK
58	2	ALRGLPIRYQTPAIR--	1	DENV2_1712_1726 57.2, DENV2_1712_1726 367.1	ALRGLPIRYQTPAIR
58	3	---RGLPIRYQTPAIRAE	3	DENV2_1714_1728 320.4	RGLPIRYQTPAIRAE
59	Consensus	KNKAIKVLRGFKKEI	-	-	-
59	1	KNGAIKVLRGFKKEI	1	DENV1_75_89 622.7	KNGAIKVLRGFKKEI
59	2	KSKAINVLRGFRKEI	1	DENV2_75_89 502.6	KSKAINVLRGFRKEI
59	3	KNKAIKILTGRKEI	1	DENV4_75_89 161.2	KNKAIKILTGRKEI
60	Consensus	ALSLTFIRSTPLVMAW	-	-	-

60	1	ALSLTFIRSTTPLVM--	1	DENV4_1277_1291 45.4	ALSLTFIRSTTPLVM
60	2	-LSLTFIRSTTPLVMA-	2	DENV4_1278_1292 105.2	LSLTFIRSTTPLVMA
60	3	--SLTFIRSTTPLIMAW	3	DENV4_1279_1293 64.7	SLTFIRSTTPLIMAW
61	Consensus	XMALVAFLRFLTIPPT	-	-	-
61	1	FMALVAFLRFLTIPP-	1	DENV2_48_62 151.7	FMALVAFLRFLTIPP
61	2	VMAFIAFLRFLAIPP-	1	DENV1_48_62 229.2, DENV1_48_62 153.3, DENV3_48_62 56.0	VMAFIAFLRFLAIPP
61	3	-MALVAFLRFLTIPPT	2	DENV2_49_63 179.2, DENV2_49_63 50.0	MALVAFLRFLTIPPT
62	Consensus	VMAFIAFLRFLAIPPTA	-	-	-
62	1	VMAFIAFLRFLAIPP--	1	DENV1_48_62 229.2, DENV1_48_62 153.3, DENV3_48_62 56.0	VMAFIAFLRFLAIPP
62	2	-MALVAFLRFLTIPPT-	2	DENV2_49_63 179.2, DENV2_49_63 50.0	MALVAFLRFLTIPPT
62	3	--AFIAFLRFLAIPPTA	3	DENV1_50_64 458.8	AFIAFLRFLAIPPTA
63	Consensus	ANCLRKNGKRVIGLS	-	-	-
63	1	TNCLRKNGKRVIGLS	1	DENV1_1850_1864 266.5, DENV1_1850_1864 61.3	TNCLRKNGKRVIGLS
63	2	ANCLRKNGKRVIGLS	1	DENV1_1850_1864 261.9	ANCLRKNGKRVIGLS
63	3	ANCLRKNGKKVIGLS	1	DENV3_1850_1864 328.5	ANCLRKNGKKVIGLS
64	Consensus	KSGAIKVLGRGFKKEI	-	-	-
64	1	KNGAIKVLGRGFKKEI	1	DENV1_75_89 622.7	KNGAIKVLGRGFKKEI
64	2	KSKAINVLRGFRKEI	1	DENV2_75_89 502.6	KSKAINVLRGFRKEI
64	3	KSGAIKVLKGFKKEI	1	DENV3_75_89 817.1	KSGAIKVLKGFKKEI
65	Consensus	IRSTMPLVMAWRTIMAVL	-	-	-
65	1	IRSTMPLVMAWRTIM---	1	DENV4_1283_1297 80.2	IRSTMPLVMAWRTIM
65	2	--STMPLVMAWRTIMAV-	3	DENV4_1285_1299 185.6	STMPLVMAWRTIMAV
65	3	---TMPLVMAWRTIMAVL	4	DENV4_1286_1300 224.0	TMPLVMAWRTIMAVL
66	Consensus	WIKETLVTFKNPHAKKQ	-	-	-
66	1	WIKETLVTFKNPHA---	1	DENV2_512_526 277.4, DENV2_512_526 195.6	WIKETLVTFKNPHA
66	2	---KETLVTFKNPHAKKQ	4	DENV2_515_529 81.0	KETLVTFKNPHAKKQ
66	3	---KKTLTVTFKNPHAKKQ	4	DENV2_515_529 158.7	KKTLTVTFKNPHAKKQ
67	Consensus	NEDHWFSRXNSXSGV	-	-	-
67	1	NEDHWFSRDNSYSGV	1	DENV3_2988_3002 372.1	NEDHWFSRDNSYSGV
67	2	NEDHWFSRGNSLSGV	1	DENV2_2988_3002 361.0	NEDHWFSRGNSLSGV
68	Consensus	KTMAMVLSIVSLFPLCL	-	-	-
68	1	KTMAMVLSIVSLFPL--	1	DENV1_1294_1308 47.3	KTMAMVLSIVSLFPL
68	2	--MAMVLSIVSLFPLCL	3	DENV1_1296_1310 61.5	MAMVLSIVSLFPLCL
69	Consensus	VXSLFMLLPTALAFH	-	-	-
69	1	VSVLFMLLPTALAFH	1	DENV1_103_117 37.3, DENV1_103_117 57.9	VSVLFMLLPTALAFH
69	2	VTMLLMLLPTALAFH	1	DENV1_103_117 87.2	VTMLLMLLPTALAFH
70	Consensus	LSLTFIKTTLSLHAW	-	-	-
70	1	LSLTFIKTTLSLHYA-	1	DENV1_1278_1292 52.7	LSLTFIKTTLSLHYA
70	2	-SLTFIKTTLSLHHAW	2	DENV1_1279_1293 24.0	SLTFIKTTLSLHHAW
71	Consensus	ISNMLNIMNRRKRSVTML	-	-	-
71	1	ISNMLNIMNRRKRSV---	1	DENV1_89_103 208.2	ISNMLNIMNRRKRSV
71	2	---MLNIMNRRKRSVTML	4	DENV1_92_106 156.5, DENV1_92_106 40.2, DENV1_92_106 288.5	MLNIMNRRKRSVTML
72	Consensus	ALGIMVLKXVRSMEKY	-	-	-
72	1	ALGIMVLKMVRSMEK-	1	DENV2_1256_1270 115.2	ALGIMVLKMVRSMEK
72	2	-LGIMVLKIVRSMEKY	2	DENV2_1257_1271 200.3	LGIMVLKIVRSMEKY
73	Consensus	GAIKVLGRGFKKEISNMLNI	-	-	-
73	1	GAIKVLGRGFKKEISN----	1	DENV1_77_91 270.8, DENV1_77_91 120.7	GAIKVLGRGFKKEISN
73	2	----VLRGFKKEISNMLNI	5	DENV1_81_95 182.1, DENV1_81_95 456.7	VLRGFKKEISNMLNI
74	Consensus	PMKLVMAFIAFLRFLAIPP	-	-	-
74	1	PMKLVMAFIAFLRFL----	1	DENV3_44_58 161.3	PMKLVMAFIAFLRFL
74	2	----VMAFIAFLRFLAIPP	5	DENV1_48_62 229.2, DENV1_48_62 153.3, DENV3_48_62 56.0	VMAFIAFLRFLAIPP
75	Consensus	ANCLRKNGKKVIGLSRKTF	-	-	-
75	1	ANCLRKNGKKVIGLS----	1	DENV3_1850_1864 328.5	ANCLRKNGKKVIGLS
75	2	----RKNKKVIGLSRKTF	5	DENV3_1854_1868 48.9, DENV2_1854_1868 675.3	RKNKKVIGLSRKTF
76	Consensus	FQRXLIFILLTAVAPS	-	-	-
76	1	FQRTLIFILLTAVAP-	1	DENV2_264_278 139.8	FQRTLIFILLTAVAP
76	2	-QRVLIFILLTAVAPS	2	DENV2_265_279 24.1	QRVLIFILLTAVAPS

77	Consensus	IKKSKAINVLRGFRKEI	-	-	-
77	1	IKKSKAINVLRGFRK---	1	DENV2_73_87 92.0	IKKSKAINVLRGFRK
77	2	---KSKAINVLRGFRKEI	3	DENV2_75_89 502.6	KSKAINVLRGFRKEI
78	Consensus	WGTIKKSKAINVLRGFRK	-	-	-
78	1	WGTIKKSKAINVLRG----	1	DENV2_70_84 88.9	WGTIKKSKAINVLRG
78	2	----IKKSKAINVLRGFRK	4	DENV2_73_87 92.0	IKKSKAINVLRGFRK
79	Consensus	LARWGTFFKKXGAIKVLKG	-	-	-
79	1	LARWGTFFKKSGAIKV----	1	DENV3_67_81 342.0	LARWGTFFKKSGAIKV
79	2	----WGTFFKLGAIKVLKG	4	DENV3_70_84 535.3	WGTFFKLGAIKVLKG
80	Consensus	NEDHWFSRXNSLSGVEGE	-	-	-
80	1	NEDHWFSRGNLSGVS----	1	DENV2_2988_3002 361.0	NEDHWFSRGNLSGVS
80	2	---HWFSRENSLSGVEGE	4	DENV1_2991_3005 263.8	HWFSRENSLSGVEGE
81	Consensus	LFTFVLLSGQITWRDMA	-	-	-
81	1	LFTFVLLSGQITWR----	1	DENV3_1165_1179 25.6	LFTFVLLSGQITWR
81	2	----FVLLSGQITWRDMA	4	DENV3_1168_1182 64.7, DENV3_1168_1182 60.0	FVLLSGQITWRDMA
82	Consensus	AGVXFVLLSGQITWR	-	-	-
82	1	AGVFFVLLSGQI----	1	DENV3_1162_1176 166.1	AGVFFVLLSGQI
82	2	----LFTFVLLSGQITWR	4	DENV3_1165_1179 25.6	LFTFVLLSGQITWR
83	Consensus	VSSVNMIXRXLNRF	-	-	-
83	1	VSSVNMISRMILNRF	1	DENV2_2724_2738 73.3	VSSVNMISRMILNRF
83	2	VSSVNMVSRLLNRF	1	DENV3_2724_2738 38.7	VSSVNMVSRLLNRF
84	Consensus	PGLFKTLTGEIGAVTLD	-	-	-
84	1	PGLFKTLTGEIGAVT---	1	DENV4_1590_1604 173.3	PGLFKTLTGEIGAVT
84	2	---LFKTLTGEIGAVTLD	3	DENV4_1592_1606 353.3	LFKTLTGEIGAVTLD
85	Consensus	WGXFKKXGAIKVLXG	-	-	-
85	1	WGTFFKLGAIKVLKG	1	DENV3_70_84 535.3	WGTFFKLGAIKVLKG
85	2	WGSFKKNGAIKVLRG	1	DENV1_70_84 736.9	WGSFKKNGAIKVLRG
86	Consensus	WGSFKKNGAIKVLRGFKK	-	-	-
86	1	WGSFKKNGAIKVLRG----	1	DENV1_70_84 736.9	WGSFKKNGAIKVLRG
86	2	----FKKNGAIKVLRGFKK	4	DENV1_73_87 117.0, DENV1_73_87 1471.8, DENV1_73_87 838.7	FKKNGAIKVLRGFKK
87	Consensus	TVGXIIIMLIPTXMAFHL	-	-	-
87	1	TVGVIIIMLIPTAMAF---	1	DENV2_102_116 61.5	TVGVIIIMLIPTAMAF
87	2	---GMIIMLIPTVMAFHL	3	DENV2_104_118 62.7	GMIIMLIPTVMAFHL
88	Consensus	AGMIIIMLIPTVXAFHL	-	-	-
88	1	AGMIIIMLIPTVVAFH-	1	DENV2_103_117 48.1	AGMIIIMLIPTVVAFH
88	2	-GMIIMLIPTVMAFHL	2	DENV2_104_118 62.7	GMIIMLIPTVMAFHL
89	Consensus	CHATFTMRLLSPVRVPNY	-	-	-
89	1	CHATFTMRLLSPVRV----	1	DENV3_1739_1753 266.0, DENV1_1739_1753 83.9, DENV2_1739_1753 30.7	CHATFTMRLLSPVRV
89	2	----TFTMRLLSPVRVPNY	4	DENV1_1742_1756 118.7	TFTMRLLSPVRVPNY
90	Consensus	DLMCHATFTMRLLSPVRV	-	-	-
90	1	DLMCHATFTMRLLSP----	1	DENV3_1736_1750 226.7	DLMCHATFTMRLLSP
90	2	----CHATFTMRLLSPVRV	4	DENV3_1739_1753 266.0, DENV1_1739_1753 83.9, DENV2_1739_1753 30.7	CHATFTMRLLSPVRV
91	Consensus	SIXNRRKXTSLCLMMM	-	-	-
91	1	SIINRRKRTSLCLMM-	1	DENV3_94_108 364.0	SIINRRKRTSLCLMM
91	2	-IVNRRKKTSLCLMMM	2	DENV3_95_109 196.5	IVNRRKKTSLCLMMM
92	Consensus	FSLXYAWKTMAMXLSI	-	-	-
92	1	FSLHYAWKTMAMVLS-	1	DENV1_1287_1301 120.5	FSLHYAWKTMAMVLS
92	2	-SLDYAWKTMAMALSI	2	DENV1_1288_1302 370.3	SLDYAWKTMAMALSI
93	Consensus	REEFXKVRSNAAIGA	-	-	-
93	1	REEFISKVRSNAAIG-	1	DENV4_2891_2905 26.9, DENV4_2891_2905 54.0	REEFISKVRSNAAIG
93	2	-EEFTRKVRSNAAIGA	2	DENV1_2892_2906 163.7	EEFTRKVRSNAAIGA
94	Consensus	GHLKCRXLMDKLXLK	-	-	-
94	1	GHLKCRCLKMDKLELK	1	DENV3_562_576 176.7	GHLKCRCLKMDKLELK
94	2	GHLKCRLRMDKLQLK	1	DENV2_562_576 317.3	GHLKCRLRMDKLQLK
95	Consensus	GLHKLGYILRDISKIPGG	-	-	-
95	1	GLHKLGYILRDISKI----	1	DENV1_3006_3020 134.1	GLHKLGYILRDISKI
95	2	----KLGYILRDISKIPGG	4	DENV1_3009_3023 125.2	KLGYILRDISKIPGG
96	Consensus	SGNIVSSVNMISRMILNRF	-	-	-
96	1	SGNIVSSVNMISRMIL----	1	DENV2_2720_2734 196.5	SGNIVSSVNMISRMIL
96	2	-----VSSVNMISRMILNRF	5	DENV2_2724_2738 73.3	VSSVNMISRMILNRF
97	Consensus	VSSVNMISRMILNRFMTK	-	-	-
97	1	VSSVNMISRMILNRF----	1	DENV2_2724_2738 73.3	VSSVNMISRMILNRF

97	2	---VNMISRMLINRFTMK	4	DENV2_2727_2741 157.7	VNMISRMLINRFTMK
98	Consensus	KEISNMLXIXNXRKX	-	-	-
98	1	KEISNMLSIIINKRKK	1	DENV3_87_101 700.0, DENV3_87_101 180.7	KEISNMLSIIINKRKK
98	2	KEISNMLNIMNRRKR	1	DENV1_87_101 980.8	KEISNMLNIMNRRKR
99	Consensus	LKKNKAIKILXGFRKEI	-	-	-
99	1	LKKNKAIKILIGFRK---	1	DENV4_73_87 73.3	LKKNKAIKILIGFRK
99	2	---KNKAIKILTGFRKEI	3	DENV4_75_89 161.2	KNKAIKILTGFRKEI
100	Consensus	SXLLKNDVPLAGPXX	-	-	-
100	1	SXLLKNDVPLAGPLI	1	DENV1_1366_1380 130.0	SXLLKNDVPLAGPLI
100	2	SALLKNDVPLAGPMV	1	DENV4_1366_1380 68.0	SALLKNDVPLAGPMV
101	Consensus	ILRHPGFTXMAAILAYTI	-	-	-
101	1	ILRHPGFTIMAAILA---	1	DENV2_242_256 407.2	ILRHPGFTIMAAILA
101	2	---HPGFTLMAAILAYTI	4	DENV2_245_259 147.3	HPGFTLMAAILAYTI
102	Consensus	LGQIMLLILCTSQILLM	-	-	-
102	1	LGQIMLLILCTSQIL---	1	DENV1_2415_2429 44.3	LGQIMLLILCTSQIL
102	2	---QIMLLILCTSQILLM	3	DENV1_2417_2431 413.9	QIMLLILCTSQILLM
103	Consensus	KEXXVTFKXPHAKXQ	-	-	-
103	1	KERMVTFKVPKAKRQ	1	DENV4_515_529 505.7, DENV4_515_529 110.0	KERMVTFKVPKAKRQ
103	2	KETLVTFKNPHAKKQ	1	DENV2_515_529 81.0	KETLVTFKNPHAKKQ
104	Consensus	FAVGLLFRRLTSREVL	-	-	-
104	1	FAVGLLFRRLTSREV---	1	DENV1_1215_1229 56.8	FAVGLLFRRLTSREV
104	2	---GLLFRRLTSREVL	4	DENV1_1218_1232 84.4	GLLFRRLTSREVL
105	Consensus	SVNMXSRLXNRFTMK	-	-	-
105	1	SVNMVSRLLLNRFTM-	1	DENV3_2726_2740 297.4, DENV3_2726_2740 66.0	SVNMVSRLLLNRFTM
105	2	-VNMISRMLINRFTMK	2	DENV2_2727_2741 157.7	VNMISRMLINRFTMK
106	Consensus	KNETWKLARASFIEVK	-	-	-
106	1	KNETWKLARASFIEV-	1	DENV1_982_996 126.1	KNETWKLARASFIEV
106	2	-NETWKLARASFIEVK	2	DENV1_983_997 380.0, DENV1_983_997 368.0	NETWKLARASFIEVK
107	Consensus	VXAFIXFLRXLIIPP	-	-	-
107	1	VMAFIAFLRFLAIPP	1	DENV1_48_62 229.2, DENV1_48_62 153.3, DENV3_48_62 56.0	VMAFIAFLRFLAIPP
107	2	VLAFITFLRVLSIPP	1	DENV4_48_62 53.5	VLAFITFLRVLSIPP
108	Consensus	MADVPLQWIASAIVLEFF	-	-	-
108	1	MADVPLQWIASAIVL---	1	DENV3_2193_2207 60.0	MADVPLQWIASAIVL
108	2	---VPLQWIASAIVLEFF	4	DENV3_2196_2210 70.7	VPLQWIASAIVLEFF
109	Consensus	MAVGXVSXLXSLLK	-	-	-
109	1	MAVGMVSILASSLLK	1	DENV2_1356_1370 158.7, DENV2_1356_1370 135.6, DENV2_1356_1370 34.4	MAVGMVSILASSLLK
109	2	MAVGLVSLLSALLK	1	DENV4_1356_1370 110.0	MAVGLVSLLSALLK
110	Consensus	MVNGVVRLLTKPWDV	-	-	-
110	1	MVNGVVRLLTKPVDV-	1	DENV2_2815_2829 153.0	MVNGVVRLLTKPWDV
110	2	-VNGVVRLLTKPWDV	2	DENV2_2816_2830 150.5	VNGVVRLLTKPWDV
111	Consensus	AIVREAIKXXRLTIL	-	-	-
111	1	AIVREAIKRLRLTI---	1	DENV3_1684_1698 29.4	AIVREAIKRLRLTI
111	2	---VREAIKKRLTIL	3	DENV1_1686_1700 136.7	VREAIKKRLTIL
112	Consensus	MAVGXVSILXSLLKN	-	-	-
112	1	MAVGMVSILASSLLK-	1	DENV2_1356_1370 158.7, DENV2_1356_1370 135.6, DENV2_1356_1370 34.4	MAVGMVSILASSLLK
112	2	-AVGIVSILLSLLKN	2	DENV1_1357_1371 76.1	AVGIVSILLSLLKN
113	Consensus	YLALXAXFKXRPXFA	-	-	-
113	1	YLALLAAFKVRPTFA	1	DENV2_1202_1216 268.0	YLALLAAFKVRPTFA
113	2	YLALMATFKMRPMFA	1	DENV1_1202_1216 357.5	YLALMATFKMRPMFA
114	Consensus	IAVGGITLFLGFTVXADM	-	-	-
114	1	IAVGGITLFLGFTVH---	1	DENV4_761_775 135.0	IAVGGITLFLGFTVH
114	2	---GGITLFLGFTVQADM	4	DENV4_764_778 132.2	GGITLFLGFTVQADM
115	Consensus	MMVLKIVRKXERYQL	-	-	-
115	1	MMVLKIVRKXERYQL-	1	DENV2_1259_1273 128.1, DENV2_1259_1273 76.4	MMVLKIVRKXERYQL
115	2	-MVLKIVRNMEKYQL	2	DENV2_1260_1274 56.3	MVLKIVRNMEKYQL
116	Consensus	RGPLKLFMALVAFLRFLTI	-	-	-
116	1	RGPLKLFMALVAFLR----	1	DENV2_42_56 43.8	RGPLKLFMALVAFLR
116	2	----KLFMALVAFLRFLTI	5	DENV2_46_60 192.3, DENV2_46_60 151.1	KLFMALVAFLRFLTI
117	Consensus	KTTFSLHYAWKTAMVLS	-	-	-

117	1	KTTFSLHYAWKTMAM---	1	DENV1_1284_1298 51.9	KTTFSLHYAWKTMAM
117	2	---FSLHYAWKTMAMVLS	4	DENV1_1287_1301 120.5	FSLHYAWKTMAMVLS
118	Consensus	PHHKILAKAIFKLTYNKV	-	-	-
118	1	PHHKILAKAIFKLT-----	1	DENV4_3054_3068 96.4	PHHKILAKAIFKLT
118	2	-----ILAKAIFKLTYNKV	5	DENV4_3058_3072 165.7	ILAKAIFKLTYNKV
119	Singleton	LMLLALIAVL TGGVT	-	DENV1_2151_2165 36.7	LMLLALIAVL TGGVT
120	Singleton	ENLKYSVIVTVHTGD	-	DENV1_414_428 23.0	ENLKYSVIVTVHTGD
121	Singleton	RAIWYMWLGARYLEF	-	DENV3_2967_2981 303.3	RAIWYMWLGARYLEF
122	Singleton	GATEIQMSSGNLLFT	-	DENV2_547_561 28.7, DENV2_547_561 48.0	GATEIQMSSGNLLFT
123	Singleton	LNTFTNMEAQLIRQM	-	DENV1_3104_3118 233.3	LNTFTNMEAQLIRQM
124	Singleton	PNYNLIVMDEAHFTD	-	DENV4_1754_1768 74.7	PNYNLIVMDEAHFTD
125	Singleton	EENLVRSVMVSAGSGE	-	DENV1_1118_1132 44.4	EENLVRSVMVSAGSGE
126	Singleton	LPVWLSYKVASAGIS	-	DENV4_2020_2034 81.1	LPVWLSYKVASAGIS
127	Singleton	SLCLMMLPATLAFH	-	DENV3_103_117 25.4	SLCLMMLPATLAFH
128	Singleton	ETLMLVALLGAMTAG	-	DENV4_2149_2163 31.3	ETLMLVALLGAMTAG
129	Singleton	QDVFCDSKLMSAAIK	-	DENV2_951_965 88.0	QDVFCDSKLMSAAIK
130	Singleton	GLVSTILASSLLRNDV	-	DENV3_1359_1373 93.3	GLVSTILASSLLRNDV
131	Singleton	GLAFSLIKNAQTPRR	-	DENV4_2481_2495 88.8	GLAFSLIKNAQTPRR
132	Singleton	GLLFSIMKNNTTSARR	-	DENV2_2481_2495 849.6	GLLFSIMKNNTTSARR
133	Singleton	RPAKNGTVMVDISRR	-	DENV1_3077_3091 31.3	RPAKNGTVMVDISRR
134	Singleton	GAGLAFSIMKSVGTG	-	DENV3_2479_2493 584.2	GAGLAFSIMKSVGTG
135	Singleton	LLTIGLSLVASVELP	-	DENV1_1231_1245 234.1	LLTIGLSLVASVELP
136	Singleton	MAHTLIMIGSNASDR	-	DENV3_1181_1195 338.3	MAHTLIMIGSNASDR
137	Singleton	ASLPTYLSSRAKLAL	-	DENV4_2107_2121 108.7	ASLPTYLSSRAKLAL
138	Singleton	KL SMGLIAIAVASGL	-	DENV4_2176_2190 25.9	KL SMGLIAIAVASGL
139	Singleton	RTGLLVISGVFPVSI	-	DENV2_1447_1461 28.1	RTGLLVISGVFPVSI
140	Singleton	KHMILVVVTTLCAII	-	DENV4_1158_1172 39.3	KHMILVVVTTLCAII
141	Singleton	GKTKRYLPAIVREAI	-	DENV2_1676_1690 405.5	GKTKRYLPAIVREAI
142	Singleton	EMILLTMKNKAWMVH	-	DENV3_476_490 82.7	EMILLTMKNKAWMVH
143	Singleton	NAWKVSCITILAAVSV	-	DENV2_1291_1305 213.0	NAWKVSCITILAAVSV
144	Singleton	ELIMKDGRVLVPCR	-	DENV2_3211_3225 40.0	ELIMKDGRVLVPCR
145	Singleton	QYKFQADSPKRLSAA	-	DENV1_807_821 30.0	QYKFQADSPKRLSAA
146	Singleton	VTRGAVLMYQGKRLE	-	DENV1_1529_1543 105.0	VTRGAVLMYQGKRLE
147	Singleton	VGGLFTSLGKAVHQQ	-	DENV4_706_720 105.3	VGGLFTSLGKAVHQQ
148	Singleton	DFGFYQVKTETTILD	-	DENV4_2258_2276 38.4	DFGFYQVKTETTILD
149	Singleton	IGGQTHLAIMAVFKM	-	DENV4_1196_1211 238.5	IGGQTHLAIMAVFKM
150	Singleton	DIELLKTEVTNPAVL	-	DENV1_323_337 395.8	DIELLKTEVTNPAVL
151	Singleton	DTLTILLKATLLAVS	-	DENV1_1440_1454 88.7	DTLTILLKATLLAVS
152	Singleton	FALLAGFMAYMIGQT	-	DENV4_248_262 67.3	FALLAGFMAYMIGQT
153	Singleton	KLTIMTGDIKGIMQA	-	DENV2_861_875 197.3	KLTIMTGDIKGIMQA
154	Singleton	TNMITLLVKLALITV	-	DENV4_1439_1453 60.2	TNMITLLVKLALITV
155	Singleton	RWSRKMLMTGT LAVF	-	DENV1_1154_1168 48.6	RWSRKMLMTGT LAVF
156	Singleton	VEVQDDGTMKIKDEE	-	DENV1_1423_1437 55.3	VEVQDDGTMKIKDEE
157	Singleton	IPLVLI IKGLNPTAI	-	DENV2_1321_1335 37.6	IPLVLI IKGLNPTAI
158	Singleton	RWLWGFLSRNKKPRI	-	DENV1_2874_2888 973.3	RWLWGFLSRNKKPRI
159	Singleton	LGMCCII TASILLWY	-	DENV2_2179_2193 30.4	LGMCCII TASILLWY
160	Singleton	ILRNPGFALLAGFMA	-	DENV4_242_256 117.3	ILRNPGFALLAGFMA
161	Singleton	GMNSRSTSLSVSLVL	-	DENV2_748_762 109.3	GMNSRSTSLSVSLVL
162	Singleton	MGVTYLAL IATFKIQ	-	DENV3_1198_1212 84.8	MGVTYLAL IATFKIQ
163	Singleton	TTLLSLTFIRTSLSL	-	DENV1_1275_1289 52.7	TTLLSLTFIRTSLSL
164	Singleton	SHHFHQLIMKDGREI	-	DENV1_3206_3220 61.9	SHHFHQLIMKDGREI
165	Singleton	LPIRYQTATKSEHT	-	DENV3_1716_1730 48.7	LPIRYQTATKSEHT
166	Singleton	LLFMILTVAANEMGL	-	DENV1_2237_2251 204.8	LLFMILTVAANEMGL
167	Singleton	WNIVKLMSGKDV FYL	-	DENV3_2616_2630 426.1	WNIVKLMSGKDV FYL
168	Singleton	ALTIKGLNPTAILLT	-	DENV2_1324_1338 791.1	ALTIKGLNPTAILLT
169	Singleton	AGQLLLMRTTWAFCE	-	DENV4_2425_2439 330.0	AGQLLLMRTTWAFCE
170	Singleton	DNQLIYVILTITII	-	DENV4_2227_2241 272.0	DNQLIYVILTITII
171	Singleton	YQNKVVRVQRPTPRG	-	DENV2_3068_3082 18.9	YQNKVVRVQRPTPRG
172	Singleton	VSGVYPMSIPATLFV	-	DENV1_1453_1467 3.8	VSGVYPMSIPATLFV
173	Singleton	ANELNYILWENNIK	-	DENV3_848_862 33.3	ANELNYILWENNIK
174	Singleton	MGKVRKDI PQWEP SK	-	DENV1_3180_3194 46.7	MGKVRKDI PQWEP SK
175	Singleton	NMEKYQLAVTIMAIL	-	DENV2_1267_1281 75.8	NMEKYQLAVTIMAIL
176	Singleton	KFWNTTIAVSMANIF	-	DENV1_2458_2472 220.0	KFWNTTIAVSMANIF
177	Singleton	CLGKAYAQMWSLMYF	-	DENV3_3249_3263 99.6	CLGKAYAQMWSLMYF
178	Singleton	SWTMKILIGVIITWI	-	DENV2_733_747 161.7	SWTMKILIGVIITWI
179	Singleton	VLMVLAHYAII GPGL	-	DENV1_2358_2372 96.7	VLMVLAHYAII GPGL

180	Singleton	KLTVVVDIIGVLEQ	-	DENV3_861_875 54.0	KLTVVVGDIIGVLEQ
181	Singleton	LSVSLVLVGVTLYL	-	DENV2_756_770 93.9	LSVSLVLVGVTLYL
182	Singleton	EDDIFRKKRLTIMDL	-	DENV2_1657_1671 146.9	EDDIFRKKRLTIMDL
183	Singleton	KCGSCVYNMMGKREK	-	DENV3_2941_2955 201.3	KCGSCVYNMMGKREK
184	Singleton	TAALLLVAHYAIIG	-	DENV2_2355_2369 133.3	TAALLLVAHYAIIG
185	Singleton	VGMVTLYLGVVMQAD	-	DENV1_763_777 74.4	VGMVTLYLGVVMQAD
186	Singleton	DLMEFIDGISLGLIL	-	DENV4_1247_1261 168.3	DLMEFIDGISLGLIL
187	Singleton	GRLITANPVVTKKEE	-	DENV3_630_644 31.7	GRLITANPVVTKKEE
188	Singleton	LNYYILLENDMKFTVV	-	DENV1_851_865 162.7	LNYYILLENDMKFTVV
189	Singleton	VQQLTKRFSGLMLQG	-	DENV2_27_41 78.1	VQQLTKRFSGLMLQG
190	Singleton	VRVQRPAKSGTVMDV	-	DENV1_3073_3087 40.0	VRVQRPAKSGTVMDV
191	Singleton	QIQPHWIAASIILEF	-	DENV2_2195_2209 99.2	QIQPHWIAASIILEF
192	Singleton	EEVMRSRWSRKMLMT	-	DENV1_1148_1162 122.7	EEVMRSRWSRKMLMT
193	Singleton	TRMVMTTTANWLWAL	-	DENV4_2865_2879 31.3	TRMVMTTTANWLWAL
194	Singleton	LETLLLLTLLATVTG	-	DENV2_2148_2162 54.0	LETLLLLTLLATVTG
195	Singleton	CLRRRVTRKHMLVV	-	DENV4_1150_1164 180.7, DENV4_1150_1164 70.0	CLRRRVTRKHMLVV
196	Singleton	YKTWAYHGSYEVKAT	-	DENV3_2795_2809 283.0	YKTWAYHGSYEVKAT
197	Singleton	LCAVQLLLMRTSWAL	-	DENV3_2423_2437 212.2	LCAVQLLLMRTSWAL
198	Singleton	AVFKMSPGYVLGVFL	-	DENV4_1207_1221 195.3	AVFKMSPGYVLGVFL

S3 Table: Selected peptides from DENV data in all the cliques and singletons

Singletons		Unique selected peptides in all cliques	
Peptide ID	Sequence	Peptide ID	Sequence
DENV1_2151_2165 36.7	LMLLALIAVLTGGVT	DENV3_1692_1706 1036.7	RRRLTLILAPTRVVA
DENV1_414_428 23.0	ENLKYSVIVTVHTGD	DENV2_50_64 1017.9	ALVAFLRFLTIPPTA
DENV3_2967_2981 303.3	RAIWYMWLGARYLEF	DENV2_12_26 353.5	TPFNMLKRERNRVST
DENV2_547_561 28.7, DENV2_547_561 48.0	GATEIQMSSGNLLFT	DENV2_1523_1537 1021.9	FHTMWHVTRGAVLMH
DENV1_3104_3118 233.3	LNTFTNMEAQLIRQM	DENV1_87_101 980.8	KEISNMLNIMNRRKR
DENV4_1754_1768 74.7	PNYNLIVMDEAHFTD	DENV4_1856_1870 1065.5	SGKKVIQLSRKTFDT
DENV1_1118_1132 44.4	EENLVRSMVSAGSGE	DENV3_1697_1711 488.7	LILAPTRVVAEMEE
DENV4_2020_2034 81.1	LPVWLSYKVASAGIS	DENV1_11_25 313.3	RPSFNMLKRVRNRVS
DENV3_103_117 25.4	SLCLMMMLPATLAFH	DENV1_268_282 478.5	IVFILLMLVTPSMAM
DENV4_2149_2163 31.3	ETLMLVALLGAMTAG	DENV3_514_528 475.0	RKELLVTFKNAHAKK
DENV2_951_965 88.0	QDVFCDSKLMSAAIK	DENV3_516_530 656.0	ELLVTFKNAHAKKQE
DENV3_1359_1373 93.3	GLVSILASSLLRNDV	DENV2_87_101 651.3	KEIGRMLNILNRRRR
DENV4_2481_2495 88.8	GLAFSLIKNAQTPRR	DENV4_3059_3073 870.7	LAKAIFKLTQNKVV
DENV2_2481_2495 849.6	GLLFSIMKNTTSARR	DENV2_52_66 579.8	VAFLRFLTIPPTAGI
DENV1_3077_3091 31.3	RPAKNGTVMDVISRR	DENV2_3260_3274 446.5	LMYFHRDLRLAANA
DENV3_2479_2493 584.2	GAGLAFSIMKSVGTG	DENV1_76_90 752.9	NGAIKVLRGFKKEIS
DENV1_1231_1245 234.1	LLTIGLSLVASVELP	DENV1_73_87 1471.8	FKKNGAIKVLRGFKK
DENV3_1181_1195 338.3	MAHTLIMIGSNASDR	DENV3_52_66 853.3	IAFLRFLAIPPTAGV
DENV4_2107_2121 108.7	ASLPTYLSSRAKLAL	DENV2_85_99 616.2	FRKEIGRMLNILNRR
DENV4_2176_2190 25.9	KLSMGLIAIVASGL	DENV1_81_95 456.7	VLRGFKKEISNMLNI
DENV2_1447_1461 28.1	RTGLLVISGVFPVSI	DENV1_85_99 407.6	FKKEISSMLNIMNRR
DENV4_1158_1172 39.3	KHMILVVVTTLCAII	DENV1_1620_1634 1645.0	EGKIVGLYGNVVT
DENV2_1676_1690 405.5	GKTKRYLPAIVREAI	DENV3_2723_2737 261.2	IVASVNMVSRLLLNR
DENV3_476_490 82.7	EMILLTMKNKAWMVH	DENV2_2720_2734 196.5	SGNIVSSVNMISRML
DENV2_1291_1305 213.0	NAWKVSCITLAAVSV	DENV1_1498_1512 151.1	GIYRILQRGLLGRSQ
DENV2_3211_3225 40.0	ELIMKDGRVLVPCR	DENV2_1200_1214 377.7	VTYLALLAAFKVRPT
DENV1_807_821 30.0	QYKFQADSPKRLSAA	DENV1_3255_3269 534.0	AQMWWQLMYFHRDLR
DENV1_1529_1543 105.0	VTRGAVLMYQGKRLE	DENV2_1833_1847 377.0	KGKTVWFVPSIKAGN
DENV4_706_720 105.3	VGGLFTSLGKAVHQV	DENV4_1217_1231 115.6	LGIFLRKLTSRETAL
DENV4_2258_2276 38.4	DFGFYQVKTEETTILD	DENV3_2726_2740 297.4	SVNMVSRLLLNRFTM
DENV4_1196_1211 238.5	IGGQTHLAIMAVFKM	DENV3_87_101 700.0	KEISNMLSIINKRKK
DENV1_323_337 395.8	DIELLKTEVTNPAVL	DENV3_1788_1802 248.6	AAIFMTATPPGTADA
DENV1_1440_1454 88.7	DTLTILLKATLLAVS	DENV4_1712_1726 412.5	ALRGLPIRYQTPAVK
DENV4_248_262 67.3	FALLAGFMAYMIGQT	DENV1_75_89 622.7	KNGAIKVLRGFKKEI
DENV2_861_875 197.3	KLTIMTGDIIKIMQA	DENV4_1278_1292 105.2	LSLTFIRSTTPLVMA
DENV4_1439_1453 60.2	TNMITLLVKLALITV	DENV1_48_62 229.2	VMAFIAFLRFLAIPP
DENV1_1154_1168 48.6	RWSRKMLMTGT LAVF	DENV1_50_64 458.8	AFIAFLRFLAIPPTA

Singletons		Unique selected peptides in all cliques	
DENV1_1423_1437 55.3	VEVQDDGTMKIKDEE	DENV3_1850_1864 328.5	ANCLRKNGKKVIQLS
DENV2_1321_1335 37.6	IPLVLIHKGLNPTAI	DENV3_75_89 817.1	KSGAIKVLKGFKKEI
DENV1_2874_2888 973.3	RWLWGFLSRNKKPRI	DENV4_1286_1300 224.0	TMPLVMAWRTIMAVL
DENV2_2179_2193 30.4	LGMCCITASILLWY	DENV2_512_526 277.4	WIQKETLVTFKNPHA
DENV4_242_256 117.3	ILRNPGFALLAGFMA	DENV3_2988_3002 372.1	NEDHWFSRDNSYSGV
DENV2_748_762 109.3	GMNSRSTSLSVSLVL	DENV1_1296_1310 61.5	MAMVLSIVSLFPLCL
DENV3_1198_1212 84.8	MGVTYLALIAITFKIQ	DENV1_103_117 87.2	VTMLLMLLPTALAFH
DENV1_1275_1289 52.7	TTLSSLTFIRTSLSL	DENV1_1278_1292 52.7	LSLTFIKTTLSLHYA
DENV1_3206_3220 61.9	SHHFHQLIMKDGREI	DENV1_92_106 288.5	MLNIMNRRKRSVTML
DENV3_1716_1730 48.7	LPIRYQTTATKSEHT	DENV2_1257_1271 200.3	LGIMVLKIVRSMEKY
DENV1_2237_2251 204.8	LLFMILTVAANEMGL	DENV2_1854_1868 675.3	RKNGKKVIQLSRKTF
DENV3_2616_2630 426.1	WNIVKLMSGKDV FYL	DENV2_264_278 139.8	FQRTLIFILLTAVAP
DENV2_1324_1338 791.1	ALTIKGLNPTAILLT	DENV2_75_89 502.6	KSKAINVLRGFRKEI
DENV4_2425_2439 330.0	AGQLLLMRTTWAFCE	DENV2_73_87 92.0	IKKSKAINVLRGFRK
DENV4_2227_2241 272.0	DNQLIYVILTILTI	DENV3_70_84 535.3	WGTFKKLGAIKVLKG
DENV2_3068_3082 18.9	YQNKVVRVQRPTPRG	DENV2_2988_3002 361.0	NEDHWFSRGNLSLGV
DENV1_1453_1467 3.8	VSGVYPMSIPATLFV	DENV3_1168_1182 64.7	FVLLLSGQITWRDMA
DENV3_848_862 33.3	ANELNYILWENNIKL	DENV3_1162_1176 166.1	AGVFFTFVLLLSGQI
DENV1_3180_3194 46.7	MGKVRKDIPQWEP SK	DENV2_2724_2738 73.3	VSSVMISRM LINRF
DENV2_1267_1281 75.8	NMEKYQLAVTIMAIL	DENV4_1592_1606 353.3	LFKTLTGEIGAVTLD
DENV1_2458_2472 220.0	KFWNTTIAVSMANIF	DENV1_70_84 736.9	WGSFKKNGAIKVLRG
DENV3_3249_3263 99.6	CLGKAYAQMWSLMYF	DENV2_104_118 62.7	GMIIMLIPTVMAFHL
DENV2_733_747 161.7	SWTMKILIGVIITWI	DENV3_1739_1753 266.0	CHATFTMRLLSPVRV
DENV1_2358_2372 96.7	VLMLVAHYAIIHG PGL	DENV3_94_108 364.0	SIINRRKRRTSLCLMM
DENV3_861_875 54.0	KLTVVVGDIIGVLEQ	DENV1_1288_1302 370.3	SLDYAWKTMAMALSI
DENV2_756_770 93.9	LSVSLVLVG VVTLYL	DENV1_2892_2906 163.7	EEFTRKVRSNAAIGA
DENV2_1657_1671 146.9	EDDIFRKKRLTIMDL	DENV2_562_576 317.3	GHLKCRLRMDKLQLK
DENV3_2941_2955 201.3	KCGSCVYNMMGKREK	DENV1_3006_3020 134.1	GLHKLGYILRDISKI
DENV2_2355_2369 133.3	TAALLLV AHYAIIG	DENV2_2727_2741 157.7	VNMISRM LINRFTMK
DENV1_763_777 74.4	VGMVTLYLGV MVQAD	DENV4_75_89 161.2	KNKAIKILTGRKEI
DENV4_1247_1261 168.3	DLMEFIDGISLGLIL	DENV1_1366_1380 130.0	SSLLKNDVPLAGPLI
DENV3_630_644 31.7	GRLITANPVVTKKEE	DENV2_242_256 407.2	ILRHPGFTIMAAILA
DENV1_851_865 162.7	LN YILLENDMKFTVV	DENV1_2417_2431 413.9	QIMLLILCTSQILLM
DENV2_27_41 78.1	VQQLTKRFSLGMLQG	DENV4_515_529 505.7	KERMVTFKVP HAKRQ
DENV1_3073_3087 40.0	VRVQRPAKSGTVMDV	DENV1_1218_1232 84.4	GLLFRRLTSREVLLL
DENV2_2195_2209 99.2	QIQPHWIAASIILEF	DENV1_983_997 368.0	NETWKLARASFIEVK
DENV1_1148_1162 122.7	EEVMRSRWSRKMLMT	DENV3_2196_2210 70.7	VPLQWIASAIVLEFF
DENV4_2865_2879 31.3	TRMVMTTTANWLWAL	DENV2_1356_1370 158.7	MAVG MVSILASSLLK
DENV2_2148_2162 54.0	LETLLLLTLLATVTG	DENV2_2815_2829 153.0	MVNGVVRLLTKPWDV
DENV4_1150_1164 180.7, DENV4_1150_1164 70.0	CLRRRVTRKH MILVV	DENV1_1686_1700 136.7	VREAIKKKLRTLILA

Singletons		Unique selected peptides in all cliques	
DENV3_2795_2809 283.0	YKTWAYHGSYEVKAT	DENV1_1202_1216 357.5	YLALMATFKMRPMFA
DENV3_2423_2437 212.2	LCAVQLLLMRTSWAL	DENV4_761_775 135.0	IAVGGITLFLGFTVH
DENV4_1207_1221 195.3	AVFKMSPGYVLGVFL	DENV2_1259_1273 128.1	MMVLKIVRKMERYQL
		DENV2_46_60 192.3	KLFMALVAFLRFLTI
		DENV1_1287_1301 120.5	FSLHYAWKTMAMVLS
		DENV4_3058_3072 165.7	ILAKAIFKLTQYQNKV

S4 Table: Analysis of overlapping clusters among different dataset

	MHC Class I	MHC Class II
T cell assays Analysis		
T cell assays+	2719	3257
T cell assays+ MHC+	2026	1260
T cell assays+ MHC-	178	175
T cell assays+ MHC (Not tested)	515	1822
MHC Analysis		
MHC+	7078	2649
MHC+ T cell assays+	2026	1260
MHC+ T cell assays-	2502	576
MHC+ T cell assays (Not tested)	2550	813
MHCLE T cell assays Analysis		
MHCLE (p)	17991	5074
MHCLE (p) T cell assays +	1124	399
MHCLE (p) T cell assays -	1554	114
MHCLE (p) T cell assays (Not tested)	15313	4561
MHCLE MHC Analysis		
MHCLE (p)	17991	5074
MHCLE (p) MHC+	2440	380
MHCLE (p) MHC-	522	54
MHCLE (p) MHC (Not tested)	15029	4640