ANTIGENIC DIVERSITY OF DENGUE VIRUS: IMPLICATIONS FOR VACCINE DESIGN

MOHAMMAD ASIF KHAN

NATIONAL UNIVERSITY OF SINGAPORE

2009

ANTIGENIC DIVERSITY OF DENGUE VIRUS: IMPLICATIONS FOR VACCINE DESIGN

MOHAMMAD ASIF KHAN
(*B. Appl. Sc. (Hons.) and M.Sc., NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

First, I thank Almighty God for His graces, guidance and for giving me the endurance to go through this strenuous exercise called PhD. I express my heartfelt gratitude to my three inspirational supervisors, Assoc/Prof Tan Tin Wee of the Department of Biochemistry, NUS, Singapore, Dr. Vladimir Brusic of Dana-Farber Cancer Institute, USA, and Professor J. Thomas August of Johns Hopkins University, USA, for their advice, guidance, continuous support and encouragement throughout the course of my candidature. I owe my sincere thanks to Dr. Olivo Miotto and Mr. Seah Seng Hong for developing the in-house computational tools used herein. I am also grateful to Dr. Srinivasan K.N., Ms. Heiny Tan, Mr. Koo Qiying, Mr. Lam Jian Hang, Dr. Zhang Guanglan, Ms. Hu Yongli, Ms. Natascha May Thevasagayam, Ms. Rashmi Sukumuran and Mr. Kenneth Lee Xunjian for their invaluable support and help during my PhD years. I am deeply indebted to Dr. Eduardo J.M. Nascimento, Dr. Kuen-Ok Jung and co-workers from the Johns Hopkins University, USA, for contributing experimental results, which validated experimentally my bioinformatics-driven research work. I am thankful to my parents, wife, siblings, and all my friends and colleagues for their continuous support, help and company over the years. I dedicate this thesis to my lovely wife Nazo. I would have not been able to complete this thesis if it was not for her continuous support, sacrifice, encouragement, and faith in me. She is truly my other half and I thank God for blessing me with her.

Table of Contents

Summary

Antigenic diversity of viruses is a significant obstacle to the development of effective therapeutic and prophylactic vaccines. Mapping T-cell epitopes among highly variable viral variants and analysing their antigenic diversity presents us with a unique opportunity to improve our understanding of immune responses to viruses and help identify peptide targets for vaccine formulation. This thesis presents a novel bioinformatics approach focusing on systematic analyses of antigenic diversity in dengue virus (DENV) sequences.

Large-scale antigenic diversity analyses presented in this thesis a) provides evidence that there are limited number of antigenic combinations in protein sequence variants of a viral species and b) suggests that a selection of short, highly conserved sequence fragments of viral proteome that also include promiscuous T-cell epitopes, applicable at the human population level, are sufficient to cover antigenic diversity of complete viral proteomes (such fragments will be referred to as *PE* for brevity).

The most important contribution of this thesis is that it provided the first, comprehensive identification and characterization of DENV *PE*s. Forty-four, highly conserved DENV *PE*s were identified and the majority was found to be immune-relevant by their correspondence to both known and putative promiscuous T-cell epitopes. Thus, these DENV *PE*s represent good targets for the development of vaccines and further experimental validation.

We defined the criteria for *PE*s, in the context of viral diversity, and developed the novel combination of bioinformatics and experimental approaches for their identification and characterization. The approach enables the design of a pipeline for large-scale systematic analysis of *PE*s within any other pathogen. The pipeline provides an experimental basis for the design of peptide-based vaccines that are

targeted to both the majority of the genetic variants of the pathogen, and the majority of human population. The generic nature and usefulness of the approach to other flaviviruses was demonstrated through the application of the pipeline to West Nile virus (WNV), which also enabled comparative analysis of characteristics of *PE*s between DENV and WNV. Such comparative analysis across pathogens of interest may provide insights into the design of better vaccine strategies.

An interesting and important finding made in this study was that there are significant differences in the conservation patterns between proteome/protein and the *PE* sites of flaviviruses, and that the patterns varied between *PE* sites, despite the flaviviruses sharing common ancestral origin, genomic architecture, and functional/structural roles of their proteins. This suggests that *PE*s may not be suitable for the formulation of a pan-*Flavivirus* vaccine and that vaccines need to be developed specific to each *Flavivirus*, preferentially using species-specific *PE*s.

This thesis provides important insights into antigenic diversity and represents a seminal contribution to the field of dengue immunoinformatics, still in its infancy. The methodology pipeline offers a paradigm shift for the field of reverse vaccinology as it enables systematic screening of all known pathogen data for *PE*s and includes multiple additional criteria for assessment of their conservation – a departure from the traditional approach where only a single or a small number of strains are studied with limited analyses of conservation.

(500 Words)

List of Figures

List of Tables

List of Abbreviations

3-D	- Three-dimensional
aa	- Amino acid
Ag	- Antigen
BLAST	- Basic local alignment search tool
BLASTP	- Protein-protein basic local alignment search tool
C	- Capsid protein
CTL	- Cytotoxic T lymphocyte
DENV/DV	- Dengue virus
DENV-1/DV1	- Dengue serotype 1
DENV-2/DV2	- Dengue serotype 2
DENV-3/DV3	- Dengue serotype 3
DENV-4/DV4	- Dengue serotype 4
DF	- Dengue fever
DHF	- Dengue hemorrhagic fever
DNA	- Deoxyribonucleic acid
DSS	- Dengue shock syndrome
E	- Envelope protein
GI	- NCBI genInfo identification number
HAV	- Hepatitis A virus
HIV	- Human immunodeficiency virus
HLA	- Human leukocyte antigen
HV	- Hantavirus
JEV	- Japanese encephalitis virus
MHC	- Major Histocompatibility Complex
NCBI	- National Center for Biotechnology Information
NS1	- Nonstructural 1 protein
NS2a	- Nonstructural 2a protein
NS2b	- Nonstructural 2b protein
NS3	- Nonstructural 3 protein
NS4a	- Nonstructural 4a protein
NS4b	- Nonstructural 4b protein
NS5	- Nonstructural 5 protein
PBMC	- Peripheral blood mononuclear cells
PE	- Highly conserved, low entropy (E) peptide (P) sequences that cover antigenic diversity
prM	- Precursor membrane protein
RNA	- Ribonucleic acid
TBEV	- Tick-borne encephalitis virus
TCR	- T-cell Receptor
WNV	- West Nile virus
YFV	- Yellow Fever virus

Chapter 1 Introduction

Viruses transmitted by blood-feeding arthropods[1] are among the most significant causes of emerging infectious diseases (Gubler, 2001). Of the approximately 130 arthropod-borne viruses (arboviruses) known to cause disease in humans (Gubler, 2001), dengue viruses (DENVs) are among the most common and medically important human pathogens (Whitehead *et al.*, 2007). DENV infection is a major health, environmental and economic problem across the globe, with most of the burden spread over the tropical and subtropical areas (Whitehead *et al.*, 2007). The virus is transmitted between humans primarily by the *Aedes aegypti* mosquito, and it causes a spectrum of manifestations ranging from an asymptomatic infection to severe disease. Disease appears most often as dengue fever (DF), while severe forms include dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS) (Whitehead *et al.*, 2007). Currently, over three billion people in more than 100 countries are at risk of dengue virus infection. Estimated 50-100 million cases of DF, and hundreds of thousands of cases of severe forms (DHF or DSS) occur annually (Whitehead *et al.*, 2007). Despite decades of effort, at present, no effective therapeutic or prophylactic vaccine exists to ease the global dengue disease burden (Whitehead *et al.*, 2007). A detailed understanding of both the virus and the human immune system will help us develop better vaccine strategies (Brusic and August, 2004).

The unique feature of DENVs compared to other flaviviruses is that they exist in nature as four genetically and immunologically distinct serotypes referred to as dengue virus serotype 1, 2, 3, and 4 (DENV-1, -2, -3 and -4) (Henchal and Putnak, 1990), each capable of causing infection in humans. Genetic differences are larger between viruses belonging to different serotypes than between viruses belonging to the same serotype. The immune responses elicited in humans after infection with a

[1] Arthropods include crustaceans, insects, arachnids, and centipedes. Crustaceans are not known as DV vectors.

DENV are abundant and directed against multiple targets within the DENV proteome (Brinton and Dispoto, 1988). Cellular immune responses, which confer protection and/or viral clearance, are an essential part of the specific immune responses to DENV infection (Kurane *et al.*, 1990; Whitehead *et al.*, 2007). The specificities of such responses are governed by major histocompatibility complex (MHC) restricted presentation of pathogen-derived peptides. Human MHC is known as human leukocyte antigen (HLA). The peptide/HLA complexes act as recognition labels, which display the contents of host cells to the surveying T cells of the immune system. Peptides that are recognized by the T cells and trigger immune responses are called T-cell epitopes. These epitopes are targets of cellular immune responses and are critical for triggering immune responses against cells infected by viruses (Hudson and Ploegh, 2002; Watts and Amigorena, 2001).

T-cell epitopes in DENV proteome are subject to changes (antigenic variation), which arise mainly from mutations and partially from recombinations of the genome (Wang *et al.*, 2002a; Wang *et al.*, 2002b). Genetic variation leading to amino acid substitution in T-cell epitopes of viruses, such as dengue (Beaumier *et al.*, 2008; Imrie *et al.*, 2007; Zivny *et al.*, 1999; Zeng *et al.*, 1996), influenza (Berkhoff *et al.*, 2007; Rimmelzwaan *et al.*, 2004; Price *et al.*, 2000; Voeten *et al.*, 2000), and HIV (Ueno *et al.*, 2007; Klenerman *et al.*, 2002; Wagner *et al.*, 1999; Harcourt *et al.*, 1998), often results in the decrease or elimination of T-cell response through reduced binding affinity of antigenic peptides to HLA molecules or T-cell receptors (Locher *et al.*, 2004). Antigenic variation enables variant viruses to escape immune recognition and prevents the build-up of specific immunity against viral variants (Haydon and Woolhouse, 1998; Sloan-Lancaster and Allen, 1996). Significant antigenic variation exists among the DENV strains, especially between serotypes (Zeng *et al.*, 1996).

Consequently, T-cell immunity to strains of one serotype is not necessarily effective against another serotype (Beaumier *et al.*, 2008), and may not even be effective against variant strains of the same serotype as has been observed with DENV antibody epitopes (Zulueta *et al.*, 2006; Blaney *et al.*, 2005; Endy *et al.*, 2004). Further, antigenic differences between DENV strains are thought to be crucial factors in complications associated with secondary DENV infections involving a serotype different from that of the primary infection. Such differences often lead to immune enhancement leading to DHF/DSS (Beaumier *et al.*, 2008; Welsh and Fujinami, 2007; Mongkolsapaya *et al.*, 2006; Mongkolsapaya *et al.*, 2003; Welsh and Rothman, 2003). Antigenic diversification of viruses, therefore, results in an increased pool of non-immune hosts with potential for severe disease symptoms. It also presents a significant obstacle for the development of therapeutic and prophylactic vaccines (Gaschen *et al.*, 2002). Mapping of T-cell epitopes across dengue variants and analysis of their antigenic diversity will improve our understanding of immune response to viral variants and help identify peptide targets for vaccine formulation.

T-cell epitopes are traditionally mapped by combination of experimental[2] methods (Sette *et al.*, 2001). A systematic analysis of a single protein involves generation of synthetic overlapping peptides spanning the whole length of the protein, followed by biochemical and functional assays of the peptides for binding to one or several HLA molecules. Binding peptides identified from these assays are then tested for recognition by T cells in functional assays. However, mapping of T-cell epitopes in DENV is not a trivial task, given the considerable sequence variation exhibited by the virus, between and within serotypes (Wang *et al.*, 2002a), as well as its propensity to frequently generate new sequences (Rico-Hesse, 1990; Trent *et al.*, 1983). Large

[2] Throughout this thesis, the term "experiment" describes procedures of "wet-lab" or laboratory bench.

number of dengue sequences is currently available in public databases (26,247 as of February 2009). These numbers render the first experimental step - synthesis and testing of a large number of peptides – impractical. Further increasing the difficulty of the task is the high polymorphism of HLA molecules (Lauemoller *et al.*, 2001). As the definition of a T-cell epitope is HLA dependent, the high polymorphism of HLA molecules in human population means that diversity of a large number of HLA specific epitopes needs to be analysed. Currently, nearly 4,600 different HLA molecules (as of April 2010) have been characterized in the human population (www.ebi.ac.uk/imgt/hla/stats.html). Taken together, the numbers of viral peptide variants and the numbers of HLA variants present an astronomical combinatorial diversity to be addressed (Brusic and August, 2004). A dismal reality of the whole mapping process is the fact that the natural prevalence of T-cell epitopes specific to a particular HLA allele in pathogen sequences is very low, approximately 0.1-5% (Brusic and Zeleznikow, 1999). This implies that from the large number of peptides tested, only a few will represent true T-cell epitopes for the specific HLA analyzed. Therefore, mapping of T-cell epitopes in DENV proteomes is analogous to the proverbial "finding needles in a haystack". Interdisciplinary approaches that combine bioinformatics, knowledge-based systems, and predictive models on one side, with biochemical and immunological approaches on the other side are essential for resolving the combinatorial complexity of DENV vaccine development.

The classification of HLA alleles into supertypes, which are groups of HLA alleles with similar peptide binding specificity (Sette and Sidney, 1999; Sette and Sidney, 1998), provides a means to help reduce the complexity arising from HLA diversity. It is estimated that the large diversity of HLA molecules in the human population can be classified into 20-30 supertypes (Brusic and August, 2004;

Doytchinova *et al.*, 2004; Lund *et al.*, 2004; Sette and Sidney, 1999). An HLA supertype includes a set of HLA molecules that typically have very similar primary sequences; they bind largely overlapping sets of peptides and, mostly belong to the same serotype. Thus, by mapping promiscuous epitopes for each of the major HLA supertypes, extensive human population coverage across different ethnic groups can be achieved (Sette and Sidney, 1999). Hence, to provide broad population coverage, it is useful to study T-cell epitopes in the context of HLA supertypes. Computational models can be used to identify candidate HLA-supertype restricted binding peptides from pathogen sequences; these peptides are potential promiscuous T-cell epitopes. Potential T-cell epitopes pre-selected by computational analysis can be rapidly validated by a small number of key experiments (Brusic *et al.*, 2004; De Groot and Rappuoli, 2004; De Groot *et al.*, 2002). Computational models in combination with experimental validation provided us a means for systematic study of antigenic conservation and variability of the large number of DENV sequences, available in public databases.

To date, mapping of T-cell epitopes in DENV and analysis of their diversity has focused on studies of a small number of common HLA molecules and, thus, only a small number of T-cell epitopes have been identified. Prior to work reported in this thesis, advanced bioinformatics tools have not been applied to mapping and diversity analysis of potential T-cell epitopes in DENV. Earlier applications have been generally limited to very simple methods, such as analysing for the occurrence of amphipathic segments, Rothbard-Taylor tetra/pentamer motifs and presence of alpha helix-preferring amino acids (Vazquez *et al.*, 2002; Kutubuddin *et al.*, 1991). These methods are of low accuracy and therefore not suitable for large-scale analysis applied in this work.

To our knowledge, none of the current DENV vaccine strategies systematically address the issue of antigenic diversity of this virus. These studies typically focus on analysis of limited strains/antigens (Rothman, 2004). Such approaches are not optimal for the development of broadly protective vaccine. An ideal dengue vaccine must be effective in providing long-lasting immunity against multiple antigenic variants of all the four DENV serotypes simultaneously and must be applicable to a large proportion of the human population. However, despite decades of work, the existing strategies have not produced such a vaccine formulation that covers the diversity of the four DENV serotypes and the human population. Though it has been recognized that a successful dengue vaccine must be tetravalent (addressing all the four serotypes), candidate DENV vaccines currently under development only consider a single variant from each serotype (Whitehead *et al.*, 2007; Rothman *et al.*, 1989). This approach has limitations because such formulations are not likely to provide a good coverage of the inter- and intra-serotype antigenic diversity of the virus. Furthermore, methods for rational selection of dengue strains and antigens, which are crucial for successful vaccination strategies, are currently not well established (Boggiano *et al.*, 2005; Duffy *et al.*, 2005; Innis and Eckels, 2003; Gaschen *et al.*, 2002). Selection of candidate strains to be included as vaccine components are mainly based on their reactogenicity[3] and immunogenicity[4] (Innis and Eckels, 2003). Such a vaccine composition may not always match the circulating strain even though they appear to be immunologically similar; this selection may limit vaccine effectiveness (Smith *et al.*, 2004). In addition, the candidate vaccine antigens are not optimized to the HLA profile of the human population since they are not selected based on possessing optimal set of targets of immune responses that are

[3] The capacity to produce adverse reactions. Least reactogenic strains are suitable for vaccine design.
[4] The capacity of an antigen to stimulate an immune response.

recognized in the context of HLA supertypes. Thus, the large diversity of the human immune system at the population level may limit the effectiveness of the vaccines developed to target certain sub-populations, or specific viral variants only.

This thesis describes original findings arising from the application of a systematic bioinformatics approach in our study of antigenic conservation and variability of DENV, and the relevance of these findings to the cellular arm of the immune system. A large-scale study of antigenic diversity of DENV was performed using a novel computational method. The author then developed a systematic bioinformatics methodology to identify and characterize peptides that cover antigenic diversity (such peptides will be referred to as *PE* for brevity) of the virus in the context of the host immune system polymorphism. The ability to encompass DENV antigenic diversity within a relatively small number of peptidic targets is important for the study of vaccine formulation targeting protection of a broad population. Future developments will require a combination of both bioinformatics and experimental approaches. The work described in this thesis presents a bioinformatics pipeline for rational selection of vaccine candidates and being generic, this method can be applied to any other pathogen. This was demonstrated through its additional applications to the West Nile virus (WNV) and other viruses, which enabled comparative analysis of characteristics of *PE*s between these pathogens. The author explored a more general conservation pattern by comparing the DENV *PE*s to corresponding sequences in 28 other viruses of the genus *Flavivirus*.

This work represents a novel contribution to the field termed "reverse vaccinology", whereby genome/proteome information is used to advance the study of vaccine formulations *in silico* in combination with targeted experimentation. Reverse vaccinology is a recently developed paradigm (Muzzi *et al.*, 2007) which uses

knowledge-based approach; it combines high-throughput screening with traditional empirical approach to vaccine development. This approach has produced several successful vaccines: such as Meningococcus-B (commercially available), Hepatitis B (commercially available), and Hepatitis C (in clinical trials). The work described in this thesis focuses on the bioinformatics component of reverse vaccinology approach. It provides important insights into understanding antigenic diversity of flaviviruses and describes newly developed methodology that enables both a global view and detailed analyses of viral proteome diversity and their implications for vaccine targeting.

1.1 Research topic

Vaccine target discovery involves studying the sequence diversity of both pathogens and human immune system to identify and characterize relevant peptides. Large amounts of sequence data produced by genomics and proteomics projects and large-scale screening of pathogen-host and antigen-host interactions are already available in public databases and are continuing to grow rapidly. The availability and growth of such large data sets are particularly relevant for vaccine target discovery as they offer the information needed for a comprehensive survey of targets and their antigenic diversity. However, experimental methods traditionally used for the study of vaccine targets are not practical for the study of large number of targets. A systematic bioinformatics approach is therefore necessary to manage and handle such large data for screening and selection of minimal set of candidate targets that can be validated by a relatively small number of key experiments. The combination of computational approaches and experimental validation, enable systematic investigation of antigen

sequences suitable as targets for vaccine formulation; this combination enables new analyses that may lead to new insights into vaccine formulations (De Groot and Rappuoli, 2004; De Groot *et al.*, 2002).

The main scope of this work focuses on the discovery of peptidic targets for vaccine formulations that cover the antigenic diversity of the four DENV serotypes and mapping these T-cell epitopes to the HLA polymorphism of human population. This thesis addresses a multi-dimensional problem arising from virus-host interactions, shown in Figure 1.1.

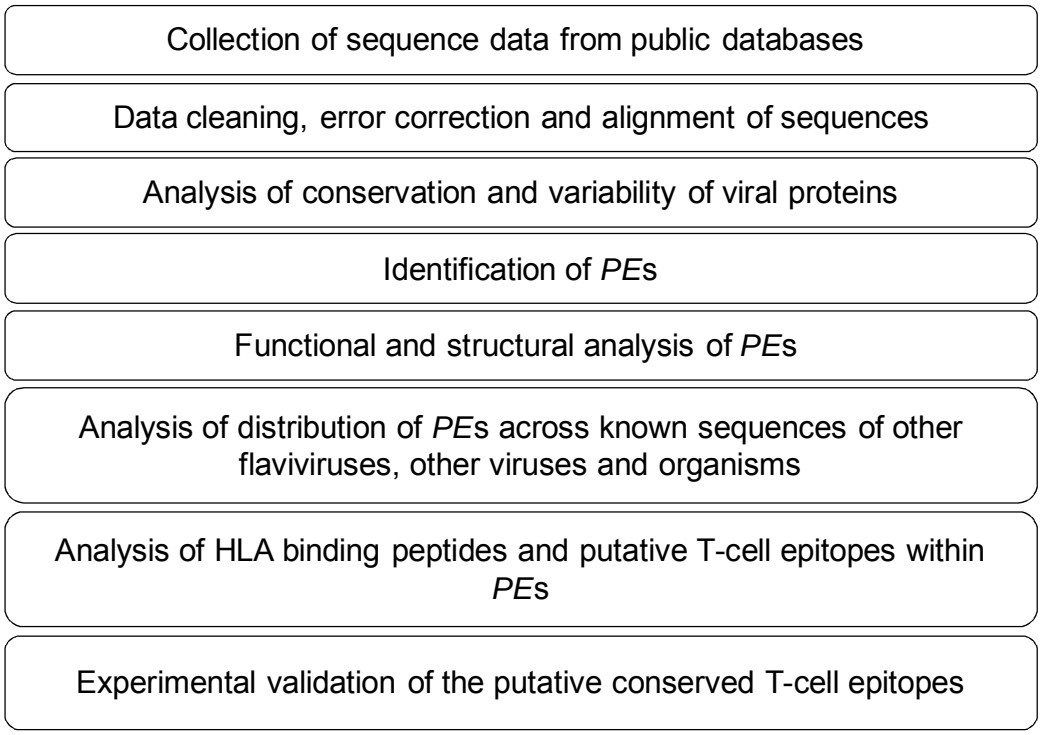


Figure 1.1: Multi-dimensional issues arising from virus-host interactions addressed in this thesis.

Additional dimensions that need to be studied, beyond the scope of this thesis, include the assessment of immunological relevance of the identified and validated conserved T-cell epitopes and assessment of their suitability for vaccine formulation.

These experimental studies can be performed in transgenic mice, or using human blood. Transgenic mice data have been shown to be relevant for understanding human immune responses (Alexander *et al.*, 2003; Tishon *et al.*, 2000; Sette *et al.*, 1994). Together, all these multiple dimensions pose an astronomical level of combinatorial complexity that can only be addressed through a combination of bioinformatics and experimental approaches.

This thesis focuses on a systematic and comprehensive computational characterization of antigenic conservation and variability of DENV. It also studies the effect of antigenic diversity to immune responses mediated by T cells, given the HLA polymorphism in human population. These results offer insights into genetic and antigenic variability in DENV and their effects to developing effective strategies for vaccine formulation against DENV infection. Both arms of the adaptive immune response (humoral and cellular) are important for protection against disease and clearance of virus. This work focuses on the cellular arm. The specific goals of this work are to develop a rational strategy for selection of dengue vaccine targets covering antigenic diversity through application of a systematic bioinformatics approach, addressing the specific issues highlighted in Figure 1.1.

1.2 Contributions

One of the main contributions of this work is the evidence that there are limited number of antigenic combinations in variant protein sequences of a viral species and that a selection of short, highly conserved sequence fragments of viral proteome that also include promiscuous T-cell epitopes, applicable at the human population level, are sufficient to cover antigenic diversity of complete viral proteomes. These insights were gained by performing a large-scale antigenic diversity analysis of DENV using a

novel *in silico* method, the specifications of which were designed and validated by the author. The method utilizes a well-defined metric that can be used with large number of sequences (either full length or partial). Importantly this method is multiple sequence alignment (MSA) independent. This makes the method robust, because MSA-based methods tend to fail when a) large number of sequences need to be aligned, b) when it is difficult to generate correct alignments because of significant sequence differences, or c) when sequences contain repeats. This method, therefore, has direct application to the analysis of any virus, in particular those that show high diversity and/or rapid evolution, such as influenza A virus and human immunodeficiency virus (HIV), which are difficult to align.

The most important, original contribution that the author provides is the first comprehensive report on identification and characterization of DENV peptides that cover antigenic diversity (*PE*s). *PE*s are short, conserved viral sequence fragments of the proteome that contain promiscuous T-cell epitopes. Forty-four (44) sequence fragments of at least nine amino acids in length were found to be highly conserved and present in ≥80% of all recorded DENV sequences (hereafter these 44 potential DENV *PEs* are referred to as pan-DENV sequences). The majority of these sequences contain putative T-cell epitopes promiscuous to multiple HLA class I and/or class II supertypes. Limited experimental validation of a number of these pan-DENV sequences proved that they contain experimentally determined promiscuous T-cell epitopes. These 44 pan-DENV sequences represent a set of potential candidate DENV vaccine targets. The observations that pan-DENV sequences have been relatively free of mutations (low peptide entropy) within the complete set of recorded sequences, with a number of them being important for viral structure and function, suggest that there is a high probability that they will remain conserved in the future. In general, the

pan-DENV sequences have relevance to multiple applications, including potential targets for prophylactic, therapeutic and diagnostic purposes.

We defined the criteria for *PEs,* in the context of viral diversity, by application of the combination of bioinformatics and experimental validation approaches described herein for the identification and characterization of immuno-relevant and highly conserved peptides. This methodology provides a novel pipeline for large-scale and systematic analysis of *PE*s of other pathogen. The bioinformatics pipeline represents the starting point for the selection of experiments that will validate vaccine targets relevant for vaccine design against multiple variants of viruses and effective for large portion of the human population. Thus, it significantly reduces effort and cost of experimentation while still providing for systematic screening. The pipeline consists of three components, namely data collection, processing and analysis. The first two components are needed to ensure that the collected data is comprehensive and "clean"[5] of errors, discrepancies and irrelevant sequences that may propagate into the subsequent analysis process shown in Figure 1.1.

The specifications for all the methods in the pipeline were designed by the author of this thesis Asif M. Khan (data collection, processing and all analyses methods, but excluding experimental validation). The author is grateful to people who contributed to this work including Dr. Olivo Miotto (contributed software tools for data collection, processing and analysis of conservation and variability), Dr. Eduardo Nascimento (experimental validation for DENV), and Dr. Kuen-Ok Jung (experimental validation for WNV). This work has been done under the supervision of Prof. J. Thomas August, Prof. Vladimir Brusic and Assoc./Prof. Tan Tin Wee. The author of this thesis alone applied the methods and tools to the study of DENV.

[5] Errors, discrepancies and irrelevant sequences were minimized to the extent possible.

The generic nature of the pipeline developed by the author was demonstrated through additional application to WNV and other viruses, such as Japanese encephalitis virus (JEV), yellow fever virus (YFV), Hepatitis A virus (HAV) and hantavirus (HV), by undergraduate students (Koo Qiying – WNV; George Au Yeung – JEV; Rashmi Sukumaran – YFV; Natascha May Thevasagayam – HAV; Hu Yongli – HV) of Dr. Tan Tin Wee's lab (Department of Biochemistry, National University of Singapore), under the supervision and with assistance from the author of this thesis. The results of the analyses enable comparative analysis for the assessment of similarities and differences in the characteristics of *PE*s across pathogens of interest, which may provide insights into the design of better vaccine strategies.

A key finding made in this study was that there are significant differences in the conservation patterns between proteome/protein and *PE* sites of flaviviruses, and that the patterns varied between *PE* sites, despite the viruses sharing common ancestral origin, genomic architecture and functional/structural role of their proteins. This is probably in response to the adaptation of each virus to the different vector-host interaction environment. This suggests that *PE*s may not be suitable for the formulation of a pan-*Flavivirus* vaccine. Instead our results indicate that vaccines need to be developed specific to each *Flavivirus*, preferentially using the species-specific *PE*s.

In summary, this work provides important insights into antigenic diversity of DENV and other flaviviruses. It represents a significant contribution to the fledgling field of dengue immunoinformatics (see Chapter 2.4). The methodology pipeline, developed as a key component of this project, brings significant advancement to the field of reverse vaccinology as it enables systematic screening of all known pathogen data for *PE*s and includes multiple additional criteria for assessment of their

conservation. This represents a departure from the traditional approach where a single strain or a small number of pathogen strains are studied with limited *in silico* analyses of conservation to identify putative antigens as vaccine targets (Vernikos, 2008; Ulmer *et al.*, 2006).

1.3 Organization of this thesis

This thesis consists of eight chapters. Chapters 1 and 2, respectively, provide an introduction to the theme and a literature review. Literature review introduces relevant readings about dengue virus, its antigenic diversity, mapping targets of immune responses in dengue viral genomes, current status and application of bioinformatics. Chapter 3 describes our large-scale antigenic diversity analysis of T-cell epitopes in DENV, while Chapter 4 reports the identification and characterization of DENV *PE*s and analysis of their potential HLA associations. These peptide sequences are potential candidates for DENV vaccine formulation. In Chapter 5, a pipeline combining systematic bioinformatics and experimental approaches for rational selection of peptide-based vaccine candidates is presented. The generic nature and usefulness of the pipeline to other flaviviruses is demonstrated in Chapter 6, coupled with comparative analysis of *PE*s between DENV and WNV. Chapter 7 describes conservation pattern of DENV *PE*s with corresponding sequences across other viruses of the genus *Flavivirus.* Original findings of the research undertaken in this thesis are summarized and discussed in Chapter 8, together with conclusions and proposed future directions.

The work presented in this thesis has been published in a series of journal articles. These include: Khan *et al*. (2006a) – the large-scale analysis of antigenic diversity of T-cell epitopes in DENV (Chapter 3); Khan *et al*. (2008) – Chapter 4,

where peptide fragments of DENV proteins that cover antigenic diversity of the four serotypes are identified and characterized; Khan *et al*. (2006b) – Chapter 5, which describes a generic, systematic bioinformatics methodology for rational selection of vaccine candidates that cover antigenic diversity; Koo, Khan *et al.* (2009a) – Chapter 6, demonstrates the generic nature and usefulness of the systematic bioinformatics approach to flaviviruses by describing its application to sequence data of WNV, and comparing the characteristics of *PE*s between DENV and WNV.

Chapter 2 Literature Review

2.1 Dengue virus (DENV)

DENVs are mosquito-borne pathogens of the family *Flaviviridae*, genus *Flavivirus,* which are phylogenetically related to other important human pathogens, such as yellow fever (YFV), Japanese encephalitis (JEV), and West Nile (WNV) viruses, among others. DENV is an enveloped, single-stranded, positive-sense RNA virus (~11 kb) that has one large open reading frame encoding a single polyprotein precursor of approximately 3,400 amino acids (~350 kDa), which is subsequently cleaved into 10 proteins by viral and host proteases: three structural (capsid, C; precursor membrane and membrane, prM/M; envelope, E) and seven nonstructural (NS) proteins (NS1, 2a, 2b, 3, 4a, 4b and 5) (Figure 2.1 and Table 2.1).

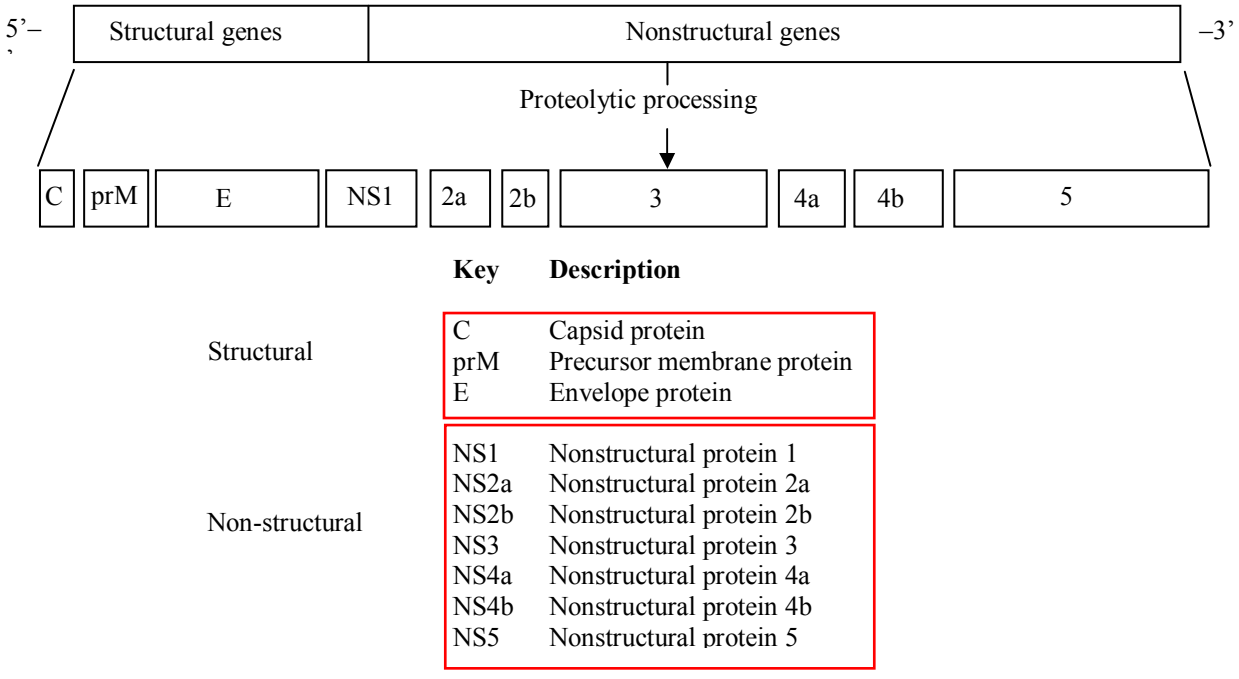


Figure 2.1: Organization of the DENV genome and proteome. A single open reading frame in the genome is translated into a single polyprotein that is cleaved by proteases to yield 10 viral proteins, of which three are structural and seven are nonstructural. [Adapted from Henchal and Putnak (1990)].

Table 2.1: Reference record for each DENV serotype from the NCBI Entrez Protein database (Benson *et al.*, 2006), providing the size in amino acids for the 10 protein products and the polyprotein.

DENV protein	NCBI accession number for each DENV serotype reference record and the size of each protein and polyprotein in amino acids			
	DENV-1	DENV-2	DENV-3	DENV-4
	AAF59976	P14340	AAM51537	AAG45437
C	114	114	114	113
prM	166	166	166	166
E	495	495	493	495
NS1	352	352	352	352
NS2a	218	218	218	218
NS2b	130	130	130	130
NS3	619	618	619	618
NS4a	150	150	150	150
NS4b	249	248	248	245
NS5	899	900	900	900
Total	3392	3391	3390	3387

Replication of viral RNA occurs in the cytoplasm in association with virus induced membrane structures and is mediated by the NS proteins. Structural proteins, enclosing the RNA, form the virus particle. The functions of the individual DENV proteins reviewed in (Lindenbach and Rice, 2003) are summarized in Table 2.2.

Table 2.2: Functions of proteins encoded by the DENV genome.

Protein	Function(s)	Reference(s)
C	Packaging of genomic RNA by forming nucleocapsid core. Mediates lipid membrane integration. May play a role in virion assembly	(Lindenbach and Rice, 2003; Markoff *et al.*, 1997)
prM	Cleavage of M from prM appears to be a crucial and terminal event in virion morphogenesis. prM may act as chaperone for folding of E protein.	(Henchal and Putnak, 1990)
E	Entry into host.	(Henchal and Putnak, 1990)
NS1	Possible role in RNA replication and pathogenesis. May be involved in assembly of the viral replicase complex and its localization to cytoplasmic membranes.	(Wallis *et al.*, 2004)
NS2a	Required for proper proteolytic processing of the C terminus of NS1.	(Falgout *et al.*, 1989)
NS2b	Forms a protease complex with NS3 which processes the viral polyprotein into separate proteins. Molecular chaperone in assisting the folding of NS3 to active conformation.	(Leung *et al.*, 2001)
NS3	Forms a protease complex with NS2b which processes the viral polyprotein into separate proteins. Implicated to play the role of RNA-dependent RNA helicase.	(Leung *et al.*, 2001; Zhang *et al.*, 1992)
NS4a & NS4b	Might be RNA replication complex cofactors along with NS5. May help anchor replicase components to cellular membranes.	(Preugschat and Strauss, 1991; Henchal and Putnak, 1990)
NS5	RNA dependent RNA polymerase	(Henchal and Putnak, 1990)

2.1.1 DENV infection in humans

DENV infection is a major mosquito-borne viral disease of humans, causing significant problem in tropical and subtropical countries. The disease ranges from asymptomatic infection, undifferentiated fever, or dengue fever (DF) to severe dengue hemorrhagic fever (DHF) with or without shock. The infection can be caused by any one of the four related, but genetically and antigenically distinct, DENV serotypes. Immunity to one serotype does not protect from infection by other serotypes (Whitehead *et al.*, 2007; Halstead, 1988). Secondary infection, caused by a serotype different from one that caused primary infection, may result in severe manifestations, such as DHF and DSS. Recent advances in our knowledge of pathogenesis and of the

immune responses elicited by DENVs emphasise the crucial role of the adaptive immune system in the control of infection (Whitehead *et al.*, 2007; Rothman, 2004). Understanding the interactions between the adaptive immune system and DENV is, therefore, important for effective strategies of vaccine development against the virus.

2.1.2 Adaptive immune responses in DENV infection

The adaptive immune response to DENV infection contributes to the resolution of the infection and has a major role in protection from re-infection. Both humoral (antibody) and cellular (T cell) components of the adaptive response are important for protection from infection and clearance of the virus (Whitehead *et al.*, 2007; Rothman, 2004; Kurane *et al.*, 1990). An ideal DENV vaccine should contain immune targets specific to both responses and for all the four serotypes. Since this study focuses on the cellular arm, antibody responses are, therefore, only briefly reviewed.

The humoral response involves antibodies produced by B-cells, which recognize both linear and conformational B-cell epitopes on the surface of DENV. Conformational neutralizing epitopes are the primary focus of DENV research on humoral responses. However, unlike linear B-cell epitopes, reliable computational tools for prediction of conformational epitopes are almost non-existent due to their complex structure (Kulkarni-Kale *et al.*, 2005).

Cellular immune responses, such as cytotoxic and helper T-lymphocyte responses, are an essential part of the specific immune responses to DENV infections (Kurane *et al.*, 1990). Cytotoxic and helper T cells, respectively, help eliminate or control viral infections by direct killing of cells infected with viruses (Bjorkman and Parham, 1990) or producing secondary signals to regulate both humoral immunity (B-cells) and cell-mediated immunity (Pulendran and Ahmed, 2006; Zinkernagel and

Hengartner, 2004; Esser *et al.*, 2003). B and T cells are known to target most of dengue viral proteins as several immunogenic[6] epitopes have been reported for each of the proteins. Immune responses to a subset of epitopes derived from an infectious pathogen can be sufficient for competent protection; thus, immune recognition of every potential epitope derived from a pathogen's proteome does not appear to be required for immune responses and protection (De Groot, 2004).

2.2 Antigenic diversity of T-cell epitopes in DENV

DENVs exist in nature as four genetically distinct serotypes. There is a considerable sequence difference between the four serotypes (Holmes and Burch, 2000). All the four serotypes are mutually distinct to the similar degree and there are suggestions that they constitute different "species" of *Flavivirus* (Kuno *et al.*, 1998). Sequence comparison studies showed 30-40% amino acid difference between serotypes (Mongkolsapaya *et al.*, 2003; Fu *et al.*, 1992). The amino acid differences within each serotype are lower but is sufficiently large to warrant the definition of clusters of DENV variants (Zhang *et al.*, 2005a; Holmes and Burch, 2000).

2.2.1 Mutation and recombination

Viral diversity across DENV genomes is a result of variation accumulated mainly through mutation (Holmes and Burch, 2000), which is partly due to the non-proofreading and, thus, error-prone nature of the viral RNA polymerase. The random mutation frequency of DENV is similar to other RNA viruses that show large

[6] Immunogenic is defined as capable of inducing an immune response; however it does not necessarily mean that this response will be useful or protective. Some immunogenic epitopes can actually enhance the disease.

diversity, such as human immunodeficiency virus (HIV) or hepatitis C virus (HCV) (Wang *et al.*, 2002a; Wang *et al.*, 2002b). Another important generator of sequence diversity for DENV is recombination, which involves exchange of genome segments between different strains (Tolou *et al.*, 2001; Uzcategui *et al.*, 2001; Holmes *et al.*, 1999; Worobey *et al.*, 1999).

The accumulation of mutation and recombination in DENV is a continuing process (Monath, 1994). There is a continuous increase in the number of newly emerging dengue variants that are unique among the members of each DENV serotype as well as between the serotypes (Rico-Hesse, 1990; Trent *et al.*, 1983). Our knowledge of the sequence diversity within each DENV serotype has risen dramatically in recent years, and the diversity is expected to further increase, recombine, and mix globally (Henchal and Putnak, 1990). The increasing sequence (genetic) diversity increases antigenic diversity because some of the changes introduced in the sequences result in changes to the T-cell epitopes through antigenic variation.

2.2.2 Antigenic variation: a challenge for vaccine design

A problem in developing a tetravalent DENV vaccine is the viral diversity (Rothman, 2004), with rather low intra-serotype, but high inter-serotype variability, resulting in both serotype-specific and serotype cross-reactive T-cell epitopes (Livingston *et al.*, 1995). This variability of related structures gives rise to a large number of variant peptide sequences with one or more amino acid differences that may function as alternative epitopes, or altered peptide ligands (Sloan-Lancaster and Allen, 1996), and affect anti-DENV host immunity (Mongkolsapaya *et al.*, 2006; Welsh and Rothman, 2003). Antigenic variation can diminish, enhance or even not affect the recognition of

viral variants by the host immune system (Takahashi *et al.*, 1989). For example, in a study by Zeng *et al*., (1996), a T-cell clone from a dengue patient tolerated a single conservative amino acid substitution from I to V at position two of the epitope peptide sequence WITDFVGKTVW (HLA-DR15 restricted), however, most other amino acid changes in this peptide abrogated the recognition. Immune escape by dengue variants often result in increased morbidity and mortality, and recurrent epidemics (Holmes and Burch, 2000; Henchal and Putnak, 1990).

In addition, immune enhancement due to cross-reactive T-cell responses may play a role in triggering deleterious immune responses, such as virus-induced immunopathology. In the case of DENV, the serotype causing secondary disease is almost always different than the serotype that induced immune response during primary infection (Rothman, 2004). Therefore, the antibodies and memory T cells induced by the primary infection typically encounter proteins containing epitopes that differ in sequence from their original targets protein. These differences may result in cross-reactive responses that contribute to the potentially fatal DSS/DHF through enhancement of the lysis of dengue virus-infected cells (Mongkolsapaya *et al.*, 2006; Welsh and Rothman, 2003). In this thesis, the author presents a method that enables selection of targets that cover a large proportion of viral sequence diversity. However, this methodology does not address the dengue virus-specific problem of protection versus immunopathology during secondary infections with a different serotype.

2.2.3 Covering antigenic diversity

Because of the significant increase of our knowledge of viral genomics and accumulated data, investigating antigenic diversity as an initial step in vaccine formulation research is necessary and prudent. Current strategies to address antigenic

diversity of virus for vaccine development include two intuitive, but contrasting approaches, (i) making use of conserved or consensus epitopes that represent multiple variants (Sette *et al*., 2001; De Groot *et al.*, 2005; Gao *et al.*, 2004), and (ii) utilizing multiple variable epitopes to represent the diverse variants, such as by using chimeric antigens containing fragments from diverse populations (Fischer *et al.*, 2007; Thomson *et al.*, 2005; Locher *et al.*, 2004), including multiple strain variants of the same antigen (Slobod *et al.*, 2005), or generating and displaying antigen diversity *in vivo* (Garcia-Quintanilla, 2007). However, none of these approaches have been explored in the field of DENV research and they do not provide insight into the relationship between genetic and antigenic diversity. Moreover, it is not clear how effective and feasible will these approaches be at circumventing the increasing future antigenic diversity in vaccine development. A systematic bioinformatics approach to analyzing antigenic diversity can aid in resolving these impending issues and provide valuable insights to help improve vaccine development strategies. Therefore, it is critical to define new methods to study antigenic diversity for vaccine development. Antigenic diversity analysis of viral antigens is an important pre-requisite to mapping T-cell epitopes.

2.3 Mapping and analyzing antigenic diversity of T-cell epitopes in DENV

2.3.1 Promiscuous T-cell epitopes: targets for mapping and analysis

Helper and cytotoxic T lymphocytes mediate cellular immune responses via the T-cell receptors (TCR) that recognize T-cell epitopes presented on cell surfaces by HLA molecules (Figure 2.2). HLA class I molecules, expressed on the surface of most nucleated cells, present endogenous epitopes, synthesized and processed in the

cytoplasm, to $CD8^+$ cytotoxic T Lymphocytes (CTLs) that eventually kill the infected cells (Shastri *et al.*, 2002; Bjorkman and Parham, 1990). $CD8^+$ cells are important in conferring immune response against intracellular viruses. On the other hand, HLA class II molecules display exogenously derived epitopes on the surface of professional antigen presenting cells (APCs), such as dendritic cells, B-cells and macrophages, for immune recognition by $CD4^+$ helper T cells. Activated helper T cells produce secondary signals for activation of both T cells and B cells (Pulendran and Ahmed, 2006; Zinkernagel and Hengartner, 2004; Esser *et al.*, 2003).

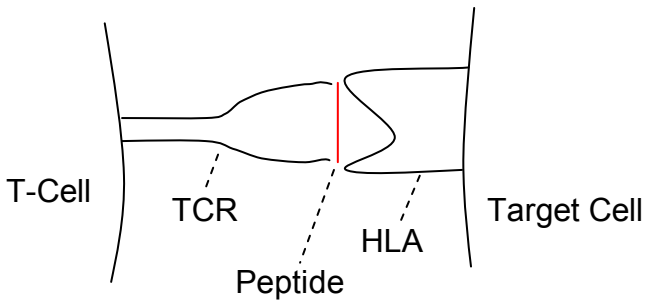


Figure 2.2: A schematic depicting the ternary complex[7] of the cellular immune arm. The complex comprises HLA class I or II molecule presenting pathogen-derived peptide, processed in the target cell, to the T-cell receptor (TCR) of the surveying T-cell of the immune system.

The recognition of peptides by the T cells are restricted by the rules/patterns governing the binding affinity and specificity of HLA molecules (Rammensee, 1995). The HLA class I groove binds to antigenic peptides of length mainly 8-11 amino acids, with nine amino acids being the typical length (Rammensee, 1995). HLA class II molecules have an open groove and bind longer peptides (12-25 amino acids in length) through a nine amino acids long core-binding region with flanking residues protruding outside of the groove (Rammensee, 1995). Some HLA class II associated peptides are reported to have multiple binding cores (Tong *et al.*, 2006).

[7] It is a term used to describe the peptide/HLA/TCR complex.

HLA genes are the most diverse of all human genes (Williams, 2001), with nearly 3,800 alleles in the human population identified to date (as of August 2009 - www.anthonynolan.org.uk/HIG). This diversity is important because it increases the spectrum of immune responses that individuals can mount, ensuring that no single pathogen can "wipe out" the entire population (Trachtenberg *et al.*, 2003). However, this diversity also increases the difficulty for the development of vaccines that will be effective across the population. The ability to trigger an effective T-cell response is partly determined by the HLA phenotype of the individual and different individuals have different HLA alleles (MacDonald *et al.*, 2001). Each individual has three to six different class I and at least that many class II HLA alleles (Brusic *et al.*, 2004; Cunha-Neto, 1999). Hence, a given vaccine may not induce identical protective immune responses in all individuals; instead, a spectrum of immune response is observed in the population (Ovsyannikova *et al.*, 2004). The discovery of similar binding specificity among different alleles of both class I and class II HLA molecules, termed as supertypes or supermotifs, provides a means to help reduce the complexity arising from HLA diversity (Sette and Sidney, 1999) and is a basis for the development of vaccines for significant population coverage. Peptides capable of binding to all or the majority of the molecules that belong to an HLA supertype are termed as "promiscuous peptides" (Brusic *et al.*, 2002) (Figure 2.3). Some peptides are also promiscuous in the context of multiple HLA supertypes, largely due to the combinatorial clustering.

The majority of HLA alleles (both class I and II) in human population can be grouped into approximately 20-30 different supertypes (Brusic and August, 2004; Doytchinova *et al.*, 2004; Lund *et al.*, 2004; Sette and Sidney, 1999). For example, the A2 supertype comprises more than 75 HLA-A2 alleles (Sidney *et al.*, 2008) of

which 14 are well studied, namely A*0201, *0202, *0203, *0204, *0205, *0206, *0207, *0208, *0209, *0210, *0211, *0212, *0213, and *0214; it should be noted that belonging to an HLA group by sequence similarity does not ensure membership within the same supertype, for example A*6802 and A*6901 are not subtypes of A2 group but they belong to the A2 supertype. Major characterised class I supertypes include A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62 (Sette and Sidney, 1999) (www.cbs.dtu.dk/services/NetCTL), while major class II supertypes include DR, [8] DQ1, DQ2, DQ3, DPw1, DPw2, DPw4, and DPw6 (Doytchinova and Flower, 2005; Southwood *et al.*, 1998). A list of major class I and II supertypes and their allelic members that are well-studied are defined in (Sidney *et al.*, 2008) and (Doytchinova and Flower, 2005), respectively.

The frequency at which the supertype alleles are expressed in various ethnicities is remarkably high (Sette and Sidney, 1999). For instance, regardless of ethnicity or gender, alleles of each of the four HLA class I supertypes (A2, A3, B44, and B7) are present in approximately 35-55% of the general population (Sette and Sidney, 1999; Sidney *et al.*, 1996), while class II DR supertype alleles alone are present in every individual, with seven most common alleles (DRB1*0101, DRB1*0401, DRB5*0101, DRB1*1501, DRB1*0701, DRB1*0901, and DRB1*1302) present in more than 80% of the general population (Southwood *et al.*, 1998). In contrast, the frequency of most individual HLA allelic forms varies significantly amongst different ethnic groups (Sette and Sidney, 1999). By combining promiscuous peptide epitopes that are specific to major HLA class I and II supertypes, extensive population coverage across different ethnic groups can be achieved. Promiscuous epitopes of a supertype are, therefore, attractive targets for the study of

[8] In this thesis, DR supertype refers to the main HLA-DR supertype defined by Southwood *et al.* (1998). At least two additional groups of HLA-DR have been proposed elsewhere.

antigenic diversity and vaccine design, because they are relevant to a large proportion of the human population.

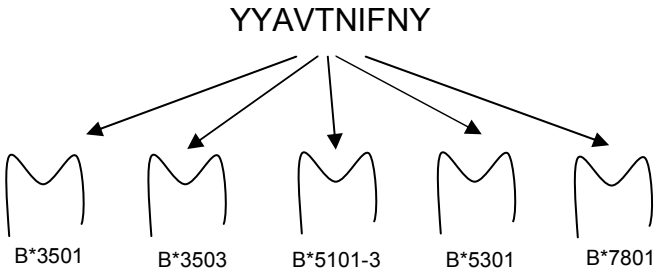


Figure 2.3: The concept of promiscuous peptides and HLA supertypes. A peptide (YYAVTNIFNY) capable of binding to multiple HLA molecules (B*3501, B*3503, B*5101-3, B*5301, B*7801) sharing similar peptide binding specificity is termed as a promiscuous peptide and the grouping of the functionally similar HLA alleles is referred to as a supertype. The HLA molecules bound by the promiscuous peptide YYAVTNIFNY are associated with the B07 supertype (Sidney *et al.*, 2008).

2.3.2 Current status of mapping and analyzing T-cell epitopes in DENV

Several studies have indicated that nearly all DENV structural and nonstructural proteins are able to elicit T-cell responses in humans (Table 2.3) (Bashyam *et al.*, 2006; Simmons *et al.*, 2005; Rothman, 2004; Brinton *et al.*, 1998; Bukowski *et al.*, 1989). The exceptions are NS2a and NS2b, for which there is still no clear evidence to suggest that they can induce cellular immune response. The structural proteins are noted as predominant source of helper T-cell epitopes, while the NS proteins mainly contain cytotoxic T-cell epitopes (Roehrig, 2003).

Table 2.3: DENV proteins reported to elicit T-cell responses in humans.

DENV proteins									
C	**prM**	**E**	**NS1**	**NS2a**	**NS2b**	**NS3**	**NS4a**	**NS4b**	**NS5**
✓	✓	✓	✓	?	?	✓	✓	✓	✓

Currently, only a limited number of human T-cell epitopes have been experimentally mapped in DENV proteins (see Table 2.4 and Appendix 1; as of March 2007). The proteins of DENV-2 are the most commonly studied, with 50 epitopes identified across all the proteins. In contrast, DENV-1 and DENV-4 are least studied, each having 15 epitopes mapped for C, E, NS3, NS4a and NS4b proteins. Minimal length T-cell epitopes have been mapped in eight of 10 structural and nonstructural DENV proteins; only NS1, NS2a and NS2b remain unmapped (though NS1 has been shown to elicit cellular responses). Among the proteins that are mapped, NS3 appears to be the best studied, with 54 epitopes mapped across the four DENV serotypes.

DENV epitopes are observed to recognize and bind an assortment of HLA alleles (Table 2.5 and Appendix 1). T-cell epitopes have been characterized for 12 different HLA alleles, with A*0201 being the most common. NS3, which contains the majority of known DENV epitopes, showed extensive HLA-restriction, covering nine of the studied 12 alleles. In contrast, characterized epitopes from the other dengue proteins are restricted by at most two alleles. The HLA-restrictions for a large number of the elucidated epitopes (38/50) were not reported. In addition, studies reporting epitopes promiscuous to multiple alleles of an HLA supertype are almost non-existent. Further, the reported T-cell epitopes have not been analysed collectively to determine their sequence diversity (antigenic diversity). Antigenic diversity was previously studied only in a small number of T-cell epitopes (Screaton and Mongkolsapaya, 2006; Welsh and Rothman, 2003).

Table 2.4: A summary of experimentally mapped DENV T-cell epitopes, their HLA-restrictions and the DENV serotype from which they were identified (DV1, 2, 3 and 4 represent DENV serotype 1, 2, 3 and 4, respectively). This data is up to date as of March 2007.

DENV protein	C				prM	E				NS1	NS2a	NS2b	NS3				NS4a				NS4b				NS5
No. of epitopes mapped	8				3	13				NA	NA	NA	54				5				8				2
No. of promiscuous epitope(s)	1				NA	NA				NA	NA	NA	3				NA				NA				NA
Serotype / No. of epitopes	DV1	DV2	DV3	DV4	DV2	DV1	DV2	DV3	DV4				DV1	DV2	DV3	DV4	DV1	DV2	DV3	DV4	DV1	DV2	DV3	DV4	DV2
	1	4	1	2	3	2	9	2	2				9	28	14	8	1	2	1	1	2	2	2	2	2

Table 2.5: HLA-restrictions of experimentally mapped DENV T-cell epitopes and the number of epitopes associated with each HLA allele.

DENV protein	HLA restricting molecule												
	A*0201	A*11	A*24	A*33	B*07	B*35	B*60	B*62	DPw2	DPw4	DR1	DRB*1501	Unknown
C	-	-	-	-	-	-	-	-	-	5	1	-	3
prM	-	-	-	-	-	-	-	-	-	-	-	-	3
E	4	-	-	-	2	-	-	-	-	-	-	-	7
NS3	-	7	3	1	2	3	1	2	2	-	-	11	24
NS4a	4	-	-	-	-	-	-	-	-	-	-	-	1
NS4b	8	-	-	-	-	-	-	-	-	-	-	-	-
NS5	-	2	-	-	-	-	-	-	-	-	-	-	-
Total	16	9	3	1	4	3	1	2	2	5	1	11	38

2.3.3 Systematic mapping and analysis of antigenic diversity of T-cell epitopes

T-cell epitopes are mapped by use of experimental methods (Sette *et al.*, 2001). A systematic analysis of a single protein involves generation of overlapping synthetic peptides spanning the whole length of the protein, followed by biochemical and functional assays of the peptides for binding to an HLA molecule (Figure 2.4). Binding peptides identified from these assays are then tested for recognition by T-cells (clones generated from human peripheral blood mononuclear cells - PBMC). The whole process is then repeated for other HLA molecules and overlapping synthetic peptides of other length to search for promiscuous epitopes.

Because peptide binding to HLA molecules is a prerequisite for T-cell recognition, testing binding of peptides of varying length to a large number of HLA molecules would require an extensive experimental effort and is the main bottleneck of the whole process. For example, a 495 amino acids long DENV envelope antigen contains 488 overlapping 8-mers (8 amino acid peptides). Testing the binding specificity of all these overlapping peptides for a single individual (generally, up to 14 HLA molecules; www.enabling.org/ia/celiac/doc/cel-hla.rtf) requires cloning, expression, and purification of the HLA molecules followed by almost 7,000 binding assays. This would then have to be repeated for overlapping peptides of other lengths.

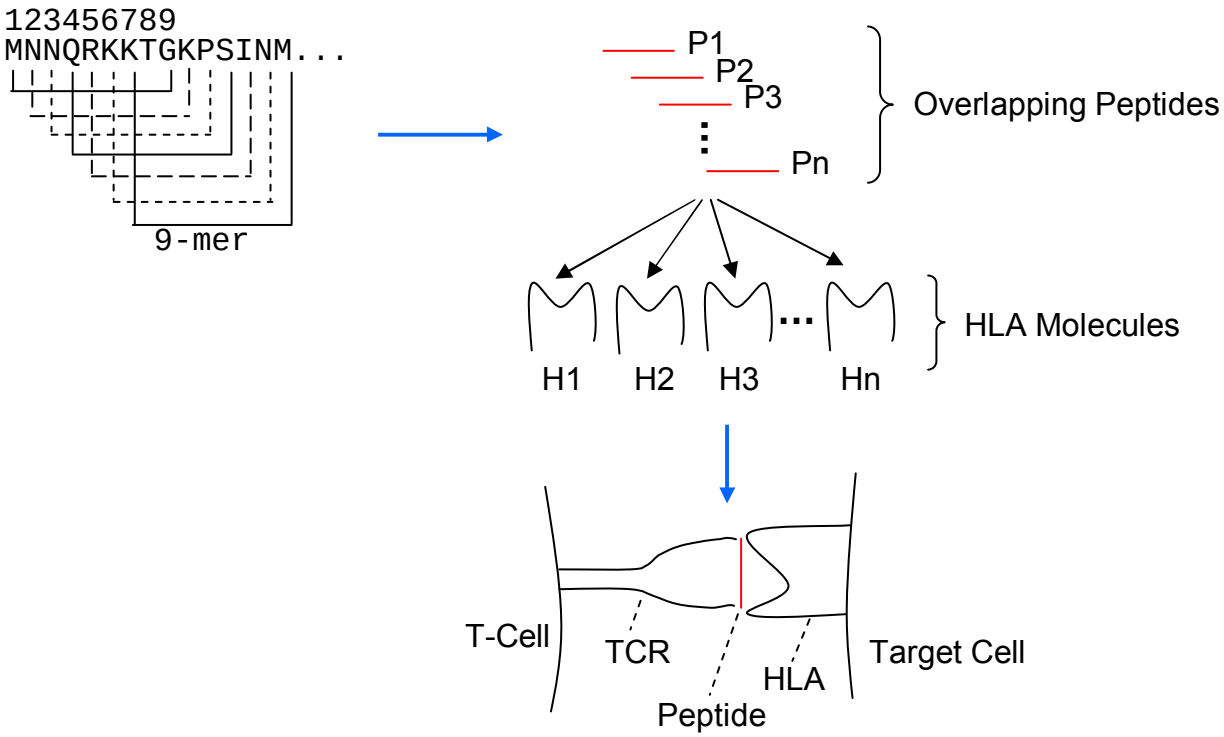


Figure 2.4: Experimental method for mapping T-cell epitopes. An all-inclusive systematic analysis of a single protein would involve the synthesis of overlapping peptides (designated by Pn, where n is the number of overlapping peptides) spanning the whole length of the protein followed by biochemical and functional assays of the peptides for binding to HLA molecules (designated by Hn, where n is the number of HLA molecules). Peptides identified from these assays are then tested for recognition by T-cells. Peptides that are recognized by the T-cells and trigger an immune response are T-cell epitopes.

Because of the cost of peptides and limited amount of human peripheral blood samples, experimental approaches are combined with a number of prediction tools developed for screening of HLA binders (Brusic *et al.*, 2004; Sylvester-Hvid *et al.*, 2002). This combination of pre-screening and targeted experimentation has dramatically accelerated the process of epitope mapping as the judicious use of the tools enable large number of unnecessary laboratory experiments to be avoided. HLA-binding peptides have been proven to minimize the time and the cost of T-cell epitope mapping (Brusic *et al.*, 2004). Highly accurate predictions can diminish discovery cost by 10-20 folds (De Groot *et al.*, 2002; Kast *et al.*, 1994).

Traditionally, computational approaches, such as sequence alignment and/or

phylogenetic programs, are used to assess the antigenic diversity of mapped T-cell epitopes. However, these approaches are not suitable for analysis of antigenic diversity because they do not analyze the sequences in the context of their interaction with the immune system. For example, instead of analyzing epitope diversity at single amino acid level, it should be analyzed according to a window size reflective of the epitope size recognized by the immune system (8-22 amino acids). Therefore, there is a need to customize the existing tools for immunological applications or develop new methods/tools for the purpose of antigenic diversity analysis.

2.4 Application of bioinformatics to analysis of viral T-cell epitopes

Mapping and analyzing the antigenic diversity of T-cell epitopes in all DENV proteins across the four serotypes is a formidable task; aside from the time-consuming and laborious nature of peptide screening, other factors also contribute making this process highly challenging. They include the large size of pathogen proteomes (including their variant strains) (Muzzi *et al.*, 2007; De Groot and Rappuoli, 2004), the great diversity of HLA molecules, (Sette *et al.*, 2001), the low (~1-5%) natural prevalence of T-cell epitopes for a given protein (Brusic and Zeleznikow, 1999), as well as restrictions imposed by high cost of peptides synthesis and limited amount of human peripheral blood available for experimentation.

Bioinformatics tools can facilitate the process of epitope mapping by identifying peptides that can potentially elicit T-cell responses. Such tools have been available for many years now – Kutubuddin *et al*. (1991) and Vazquez *et al*. (2002) utilized basic bioinformatics methods (Rothbard and Taylor's pattern method and amphipathic α-helix formation) to predict potential T-cell epitopes in the DENV envelope and prM proteins, respectively, prior to experimental validation. However,

bioinformatics did not feature prominently in earlier DENV research and its incorporation into wet lab procedures appears to be a recent trend.

Several papers describing the application of advanced prediction tools to study DENV epitopes have been published in the recent years. For example, Bashyam *et al.* (2006) used an advanced epitope prediction algorithm (BIMAS) to generate a series of nonamers that were predicted to possess the HLA-A*0201 binding motif. Sanchez *et al.* (Sanchez *et al.*, 2006) also used the BIMAS algorithm to predict B*60-, A*24- and A*2-binding peptides. Mazumder *et al*. (Mazumder *et al.*, 2007) used NetMHC to predict HLA class I binding peptides and MHCPred and RANKPEP to predict Class II binders from the E protein. However, applications of these tools at large-scale to predict and analyze all available DENV sequence data in public databases was not performed.

Limited application of bioinformatics in mapping and analyzing DENV epitopes in the past may be attributed to the lack of the desired accuracy of the prediction tools. Accumulation of experimental data coupled with ongoing efforts to improve software performance based on the understanding of complex molecular events involved in antigen processing and presentation has led to greater reliability (85-95% accuracy) of the recent advanced tools and web-servers (Lin *et al.*, 2008; Wang *et al.*, 2008). Sophisticated prediction methods for major steps of HLA class I and II antigen processing and presentation pathways have been developed, reviewed in (Brusic *et al.*, 2004). The list of servers for prediction of antigen processing and HLA binding are provided in Table 2.6. Prediction tools are also being directly linked to information derived from microarray, proteomics, and other technologies to provide more information on the potential antigens (Grandi, 2003; De Groot and Rothman, 1999). In addition, efforts are underway to standardize peptide binding

affinity assays, which can greatly facilitate utilization of the growing bank of HLA-peptide information (Petrovsky *et al.*, 2003; Stenman, 2001). Therefore, there is an urgent need for the DENV-human immunome[9] research field to integrate bioinformatics (specifically immunoinformatics) as a key part of the experimental strategy as it has the potential to revolutionise the way researcher study the interactions between the virus and host immune response for therapeutic and/or prophylactic purposes.

[9] The immunome is defined as the full set of pathogen's antigens interacting with the host immune system

Table 2.6: An overview of bioinformatics prediction servers for mapping putative T-cell epitopes.

Prediction parameters	Prediction tool	URL/Reference	Remarks
Antigen processing: Proteasome cleavage	MAPPP Proteasome cleavage prediction	http://www.mpiib-berlin.mpg.de/MAPPP/cleavage.html	-
	NetChop	http://www.cbs.dtu.dk/services/NetChop/	
	PAProC	http://www.paproc.de/	
	Pcleavage	http://www.imtech.res.in/raghava/pcleavage/	
Antigen processing: TAP binding	$PRED^{TAP}$	(Zhang *et al.*, 2006)	-
	SVMTAP	http://www-bs.informatik.uni-tuebingen.de/Services/SVMTAP/	
	TAPPred	http://www.imtech.res.in/raghava/tappred/	
Antigen processing: HLA binding	BIMAS	http://www-bimas.cit.nih.gov/molbio/hla_bind/	HLA Class I, single allele
	HLA Epitope binding prediction	http://hlaligand.ouhsc.edu/prediction.htm	
	IEDB Analysis Resource MHC-I binding predictions	http://tools.immuneepitope.org/analyze/html/mhc_binding.html	HLA Class I, promiscuous epitopes
	MAPPP MHC-I binding prediction	http://www.mpiib-berlin.mpg.de/MAPPP/binding.html	
	MMBPred	http://www.imtech.res.in/raghava/mmbpred/	HLA Class I, promiscuous epitopes, prediction of mutated HLA binders
	NetMHC	http://www.cbs.dtu.dk/services/NetMHC	HLA Class I, single allele
	PREDEP	http://margalit.huji.ac.il/	
	SMM	http://zlab.bu.edu/SMM/	
	FDR4	http://www.imtech.res.in/raghava/fdr4/submit.html	HLA Class II, single allele
	HLA-DR4Pred	http://www.imtech.res.in/raghava/hladr4pred/	
	IEDB Analysis Resource MHC Binding Prediction	http://tools.immuneepitope.org/tools/matrix/iedb_input?matrixClass=II	
	MOT	http://www.imtech.res.in/raghava/mhc/page4.html	
	MHC2Pred	http://www.imtech.res.in/raghava/mhc2pred/	HLA Class II, promiscuous epitopes
	ProPred	http://www.imtech.res.in/raghava/propred/	
	TEPITOPE	(Bian and Hammer, 2004)	

Prediction parameters	Prediction tool	URL/Reference	Remarks
	ARB Matrix	http://epitope.liai.org:8080/tools/matrix/iedb_input?matrixClass=I,II	HLA Class I & II, single allele
	MHC binder prediction	http://www.vaccinedesign.com/	HLA Class I & II, promiscuous epitopes, selects best predictions from five algorithms
	HLAPred	http://www.imtech.res.in/raghava/hlapred/	HLA Class I & II, promiscuous epitopes
	MHC-BPS	http://bidd.cz3.nus.edu.sg/mhc/	HLA Class I & II, single allele
	MHCPred	http://www.jenner.ac.uk/MHCPred	
	MULTIPRED	(Zhang *et al.*, 2005b)	HLA Class I & II, promiscuous epitopes, predicts hotspots
	SVMHC	http://www-bs.informatik.uni-tuebingen.de/Services/SVMHC	HLA Class I & II, single allele
	SVRMHC prediction server	http://svrmhc.umn.edu/SVRMHCdb/	
	SYFPEITHI	http://www.syfpeithi.de/Scripts/MHCServer.dll/EpitopePrediction.htm	HLA Class I & II, promiscuous epitopes
Antigen processing: Integrating HLA class I binding, TAP binding, and proteasomal cleavage predictions	EpiJen	http://www.jenner.ac.uk/EpiJen/	Single HLA allele
	MHC-I ligand processing predictions	http://tools.immuneepitope.org/analyze/html/mhc_processing.html	Promiscuous epitopes
	MHC-pathway	http://www.mhc-pathway.net/	Single HLA allele
	NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	Promiscuous epitopes
	WAPP	http://www-bs.informatik.uni-tuebingen.de/WAPP	Single HLA allele
	nHLAPred	http://www.imtech.res.in/raghava/nhlapred/	Promiscuous epitopes, does not predict TAP binding
	PEPVAC	http://bio.dfci.harvard.edu/PEPVAC/	
	ProPred-I	http://www.imtech.res.in/raghava/propred1/	
	RankPep	http://bio.dfci.harvard.edu/Tools/rankpep.html	
Patterns of T-cell epitopes	CTLPred	http://www.imtech.res.in/raghava/ctlpred/	-
Immunological hotspots (clusters of promiscuous epitopes)	Hotspot Hunter	(Zhang *et al.*, 2008)	-

2.5 Chapter summary

- Analysis of antigenic diversity of T-cell epitopes is necessary to develop strategies to cover the diversity and aid in rational selection of vaccine targets. Antigenic diversity analysis of viral antigens is an important pre-requisite to mapping T-cell epitopes.
- The discovery of similar binding specificity among different alleles of both class I and class II HLA molecules, termed as supertypes or supermotifs, helps reduce the diversity of the HLA molecules and provides a basis for the development of vaccines effective at the population level. Therefore, T-cell epitopes that are promiscuous to multiple alleles of an HLA supertype are attractive targets to map and analyse for antigenic diversity.
- A limited number of human T-cell epitopes have been experimentally mapped in DENV proteins to date.
- Experimental approaches are combined with a number of prediction tools to accelerate the process of epitope mapping as the judicious use of the tools enable large number of laboratory experiments to be avoided. Highly accurate predictions can diminish discovery cost by 10-20 folds. However, mapping putative T-cell epitopes in dengue proteins using bioinformatics tools is hardly existent.
- A systematic bioinformatics-based approach to proteome-wide mapping and analysis of potential DENV T-cell epitopes can provide important insights into covering antigenic diversity and improving the efficacy of research by assisting in selection of critical experiments for vaccine formulation.

Chapter 3 Large-scale Analysis of Antigenic Diversity of T-Cell Epitopes in Dengue Virus

3.1 Introduction

While there is a correspondence between genetic and antigenic evolution of viruses, genetic changes can result in disproportionately large antigenic changes (Smith *et al.*, 2004; Morvan *et al.*, 1990). Though genetic and antigenic diversity in DV strains is evident (Rico-Hesse, 2003), large-scale and detailed systematic analyses that explore their relationship have not been reported. Earlier studies of genetic diversity focused on clade diversity and replacement (Zhang *et al.*, 2005a), mutation spectra (Chao *et al.*, 2005), conserved regions (Schein *et al.*, 2005) and implications for clinical manifestations (Holmes and Burch, 2000). While studies of antigenic diversity (diversity of targets of immune responses in protein sequences) focused on experimental mapping of limited T-cell epitopes (Simmons *et al.*, 2005; Mongkolsapaya *et al.*, 2003; Loke *et al.*, 2001; Kurane *et al.*, 1998; Gagnon *et al.*, 1996) and subsequent analysis of their diversity. Understanding this relationship between genetic and antigenic diversity is important for the study of vaccine development, especially in rapidly mutating viruses. In the study reported in this chapter, the author focuses on protein sequence diversity, and thus only considers genetic variations (non-synonymous mutations) that affect the protein sequences.

We developed a bioinformatics method to analyze antigenic diversity in the context of T-cell mediated immune responses (Khan *et al.*, 2006a). Antigenic diversity of more than 9000 DV protein sequences reported in the NCBI Entrez Protein database (Wheeler *et al.*, 2005) were studied. The study aimed to identify a minimal set of sequences that encodes the complete antigenic diversity of short peptides from all known sequences of DV serotypes. Short peptides, principally 9-mers were studied because they represent the predominant length of binding cores of T-cell epitopes. The relationship between short-peptide antigenic diversity and protein

sequence diversity of DV was analysed; the analysis was performed at two time points (reported data up to June 2004 and up to December 2005) to help understand the effects of the accumulation of sequence data to the relationship. The effects of sequence determinants on antigenic diversity of short peptides were also assessed. This study provided a framework for large-scale, systematic analysis of antigenic diversity for the protein sequences of any virus.

The author would like to declare here that the method developed for antigenic diversity analysis and the results for the up to June 2004 time point data set are part of his MSc thesis (Khan, 2005) with the National University of Singapore. In consultation with his supervisors, the author determined that it was necessary to update the study, because the availability of new DV sequences in public databases (157% increase) would (i) produce a statistically more accurate representation of antigenic diversity of T-cell epitopes in DV, ii) provide additional validation of the method using independent data set, and iii) help assess how increases in sequences affects the relationship between antigenic and genetic diversity.

3.2 Materials and methods

3.2.1 Dengue virus data collection

All DV protein sequence entries present in the NCBI Entrez Protein database (Brown *et al.*, 2003) were collected in June 2004 (comprising all reported data up to June 2004) and then again in December 2005 (comprising all reported data up to December 2005). Data retrieval was performed through the NCBI taxonomy browser (Wheeler *et al.*, 2005) and the respective taxonomy ID for each of the dengue serotypes (DV1-

4) are 11053, 11060, 11069 and 11070. The collected entries for both time points were processed separately using identical procedures.

3.2.2 Data processing: cleaning and grouping

Data compiled from public databases are prone to errors and discrepancies (Srinivasan *et al.*, 2002), which if not corrected may obscure the results of analyses. Therefore, the author inspected the collected DV entries and corrected errors and discrepancies. Individual protein sequences were extracted from collected entries for each DV serotype and grouped according to the 10 dengue proteins for analysis. The extraction and grouping was facilitated by creating a searchable database (McGinnis and Madden, 2004) of all collected DV sequences and performing blast (parameters: filter – no; expect – 100; descriptions & alignments – 20,000) against the database by use of a sample sequence for each DV protein. The sample sequences were obtained from the NCBI Entrez Protein database (Wheeler *et al.*, 2005) reference record for each dengue serotype (DV1: AAF59976; DV2: P14340; DV3: AAM51537; DV4: AAG45437). A total of 40 blast searches were performed (4 serotypes x 10 proteins of each serotype) and the blast hits obtained for each search were all the data available for the respective protein in the collected DV records. The cleavage site information for each dengue protein was used to assess the reliability of the blast results and filter out irrelevant hits. The cleavage site information was obtained from the annotation of the NCBI Entrez Protein database reference records and the literature (Osatomi and Sumiyoshi, 1990).

Duplicate or identical sequences were then removed from the resulting datasets of each DV protein and the unique sequences were retained for further

analysis. Both full length and partial unique sequences of each dengue serotype protein were used in all the analysis, unless indicated otherwise in the sections below.

3.2.3 Extent of amino acid variation within and across DV serotype proteins

Pairwise percentage amino acid identity for the full length unique sequences of each dengue protein, intra- and inter-serotype, was computed by use of ClustalW 1.83 (Thompson *et al.*, 1994) with default parameters. This was done to survey the extent of amino acid variation in the latest, comprehensive dengue dataset of 2005.

3.2.4 Protein sequence and antigenic diversity analysis of DV

In this study, protein sequence diversity of a dengue protein was defined as the total number of unique sequences reported in the dataset for the protein. Sequences having at least a single amino acid difference between them were considered as unique.

Antigenic diversity of a dengue protein was defined in this study as the minimal set of unique sequences required to represent the complete set of overlapping 9-mer peptides encoded by all unique sequences reported in the protein dataset. A bioinformatics method that performs exhaustive search to determine the minimal set for a given protein was developed. The method comprises of two steps: (a) generation of a set of overlapping 9-mers from the entire length of all unique sequences reported in the protein dataset, followed by (b) identification of a minimal set of unique sequences that represents all the unique 9-mers. The union of such sets for all the 10 proteins of a dengue serotype represents the antigenic diversity of the proteins for the serotype. The specifications for the method were defined and validated by the author

and the computer program for the method was written in Perl and C languages by Seah Seng Hong, a programmer in the supervisor's lab.

In the first step of the method, overlapping 9-mers from the entire length of each unique sequence are generated because the whole length is assumed to contain T-cell epitopes. This assumption was based, firstly, on the estimate that from a complete set of overlapping peptides (9 or 10-mers) spanning a protein, on average, 0.1-5% of the peptides will bind to any particular HLA molecule (Brusic and Zeleznikow, 1999). Secondly, given the large number of HLA molecules (more than 2532 known as of September 2006; www.ebi.ac.uk/imgt/hla/stats.html), the vast majority of the complete set of overlapping peptides are highly likely to bind at least one molecule from the total HLA pool. Thus, each overlapping peptide is a potential T-cell epitope. Further, this assumption ensured the coverage of all possible candidate 9-mer T-cell epitopes that can be present across the entire length of the unique sequence. Antigenic diversity study was focused on 9-mers because they represent the predominant length of HLA class I T-cell epitopes, as well as the binding core of HLA class II T-cell epitopes (Rammensee, 1995). Furthermore, we performed a preliminary analysis using 8-mers and 10-mers, which did not produce notably different results compared to the analysis of 9-mers (data not shown). A small number of 9-mers derived from the unique sequences contained unknown residues (denoted by "X") and, hence, were excluded from the analysis because they were antigenically non-informative.

3.2.5 Determining the effects of sequence determinants on antigenic diversity

The effects of two sequence determinants were studied on antigenic diversity: i) the number of viral sequences in the studied set and ii) the length of protein antigens. The

study was performed on unique sequences of the DV2 envelope protein (retrieved in 2005) because it provided a sufficiently large and well-defined dataset (198 full length sequences). Test datasets with different numbers of sequences (20, 40, 60, 80, 100, 120 and 140 sequences) and different lengths (23, 46, 128, 138, 276 and 460 aa) were randomly derived from the envelope dataset with repeated sampling (20 repeats). Any duplicate sequences were removed from the test datasets. The minimal set of sequences that represents the complete short-peptide antigenic diversity was determined for each test dataset. These minimal sets were used to analyze the effects of the sequence determinants on antigenic diversity.

3.3 Results

3.3.1 DV serotype protein datasets

Data of June 2004 (Table 3.1), collected from the NCBI Entrez Protein database, contained a total of 3699 sequences representing the 10 proteins encoded by the genomes of the four serotypes (Table 2.1 and Figure 2.1). The number of these reported sequences increased nearly three-fold during the following 18 months (9512 sequences; see Table 3.1). The removal of duplicates (identical protein sequences) reduced these collected sequences to 1318 (2004) and 2419 (2005) unique sequences (Table 3.1). More than 64% of the sequences collected in 2004 were identical and, thus, redundant, and the redundancy increased by approximately 10% in 2005 (to 75%). The number of reported unique sequences varied greatly among the proteins, ranging from 69 NS4a to 998 E sequences in the 2005 set (Table 3.2). Minor errors of annotation, mainly of the cleavage sites, were identified for 17 sequences (Appendix 2) and the source databases of these sequences, such as NCBI and Swiss-Prot, were

notified for correction. An offshoot development of this, beyond the scope of this thesis, was that the reporting of the DV annotation errors by the author of this thesis to the personnel (Philippe Le Mercier *et al*.,) at the Swiss-Prot database made them later realize that it was not an isolated case specific to DENV; it was a general problem for sequence records in public databases of many other viruses. This observation in particular, among other reasons, made them start an international initiative to correct such problems across public databases by developing common standards for virus sequence records, such as a standard nomenclature to name virus isolates. Towards this initiative, the author contributed in proposing the nomenclature specifications.

Table 3.1: Number of collected and unique protein sequences for each dengue serotype as of 2004 and 2005 and the corresponding increase in data between the two time points.

Dengue serotype	Data retrieved in 2004 (#)		Data retrieved in 2005 (#)		Increase (#)	
	Collected sequences	Unique sequences	Collected sequences	Unique sequences	Collected sequences	Unique sequences
DV1	744	359	2318	724	1574	365
DV2	1426	507	3351	697	1925	190
DV3	597	230	2520	678	1923	448
DV4	932	222	1323	320	391	98
Total	*3699*	*1318*	*9512*	*2419*	*5813*	*1101*

Table 3.2: Total number of unique sequences for the proteins of the four DV serotypes, as of 2004 and 2005.

Protein	No. of unique sequences (all four serotypes)	
	2004	2005
C	107	196
prM	126	220
E	495	998
NS1	150	224
NS2a	95	142
NS2b	59	78
NS3	80	164
NS4a	37	69
NS4b	57	88
NS5	112	240
Total	*1318*	*2419*

3.3.2 Intra- and inter-serotype amino acid sequence variability of DV proteins

Earlier studies of dengue proteins, mainly E and NS1 (Holmes and Twiddy, 2003; Twiddy *et al.*, 2003; Twiddy *et al.*, 2002; Fu *et al.*, 1992; Rico-Hesse, 1990), have shown substantial amino acid sequence diversity both within and between the serotypes. In this study, the extent of amino acid variation among DVs was surveyed by calculating pairwise percentage amino acid identity of unique sequences for each dengue protein, intra- and inter-serotype, using the large dengue data set of 2005. The intra- and inter-serotype percentage sequence identities (PSI) for all dengue proteins are shown in Table 3.3.

Table 3.3: Minimum and maximum percentage sequence identity range for each dengue protein, intra- and inter-serotype. The average percentage sequence identities (PSI) are shown for inter-serotype comparisons.

		DV1	DV2	DV3	DV4	Average PSI (%)
C	DV1	88-99				65
	DV2	56-75	81-99			
	DV3	75-84	53-66	91-99		
	DV4	61-68	57-69	54-60	94-99	
prM	DV1	92-99				68
	DV2	62-75	79-99			
	DV3	75-82	60-72	93-99		
	DV4	62-67	60-71	64-70	96-99	
E	DV1	89-99				65
	DV2	58-70	80-99			
	DV3	72-79	60-69	92-99		
	DV4	58-66	55-65	61-64	94-99	
NS1	DV1	93-99				72
	DV2	68-75	85-99			
	DV3	77-80	69-75	94-99		
	DV4	67-70	68-73	70-74	93-99	
NS2a	DV1	90-99				39
	DV2	36-40	93-99			
	DV3	43-48	35-40	93-99		
	DV4	35-39	33-36	36-41	89-99	
NS2b	DV1	93-99				60
	DV2	56-62	95-99			
	DV3	66-70	58-63	96-99		
	DV4	56-62	54-59	56-59	94-99	
NS3	DV1	97-99				79
	DV2	78-80	96-99			
	DV3	84-86	79-81	97-99		
	DV4	75-77	75-77	77-79	97-99	
NS4a	DV1	92-99				60
	DV2	56-61	96-99			
	DV3	63-68	56-63	92-99		
	DV4	56-60	59-64	56-62	94-99	
NS4b	DV1	95-99				78
	DV2	75-79	95-99			
	DV3	81-85	75-79	97-99		
	DV4	75-78	77-81	76-79	97-99	
NS5	DV1	96-99				77
	DV2	77-79	95-99			
	DV3	80-82	77-79	96-99		
	DV4	73-76	72-75	74-77	95-99	

The intra-serotype percentage sequence identity was between 92% and 99%, except for C, prM, E and NS1 of DV2, which showed minimum sequence identities ranging from 79% to 89%. In contrast, the average inter-serotype percentage sequence identity was in the range of 60-79%, except for NS2a. The NS3, NS4b and NS5 proteins are highly conserved across the serotypes, with average sequence identities in the range of 77-79%, probably because of their involvement in forming the RNA replication complex (Preugschat and Strauss, 1991). The NS2a protein is the most diverse across the serotypes (average PSI of 39%), though it is highly conserved within each serotype. The inter-serotype diversity observed for NS2a is comparable to

the inter-*Flavivirus* diversity of the envelope protein, which shows approximately 40% amino acid identity (Mukhopadhyay *et al.*, 2005).

3.3.3 Minimal sequence sets representing DV antigenic diversity

In addition to identical protein sequences, another source of sequence redundancy, relevant to this study, is the presence of antigenically redundant sequences. These sequences exist because of the identity of many amino acid residues among the individually unique protein sequences, resulting in the presence of T-cell epitopes that are identical among viral variants. Antigenically redundant sequences can be removed without loss of information on antigenic diversity among the sequences in a dataset. For example, in a dataset of three sequences, if all the overlapping 9-mers in one sequence have a match in at least one of the other two sequences, the antigenic diversity of this sequence can be covered by the other two sequences combined, thus rendering the first sequence antigenically redundant (Figure 3.1).

After the removal of duplicate sequences, the removal of antigenically redundant sequences using the bioinformatics method resulted in a further reduction of the number of dengue unique sequences to a total of 969 (2004 set) or 1684 (2005 set). These two sets represent the complete antigenic diversity of short peptides for all four dengue serotypes (Table 3.4). The increase in the number of unique sequences required to represent the complete antigenic diversity of short peptides in the four dengue serotypes in 2005, compared to 2004, is an indication that more short-peptide antigenic diversity was found in the new sequences accumulated in the database. However, the percentage of unique sequences required to represent the complete short-peptide antigenic diversity of all four dengue serotypes in 2005 decreased (from 74% in 2004 to 70% in 2005) because of an increase in antigenic redundancy. This

observation indicates that the increase in the number of unique protein sequences (representing protein sequence diversity) deposited in public databases is generally accompanied by a slower increase in short-peptide antigenic diversity.

A) Three unique sequences from the NCBI Entrez Protein database.

```
               1854039  ASIILEFFLMVLLIPEPDRQRT
              17129648  ASIILEFFLMVLLIPEPDRLRT
              37963458  ASIILEFLLMVLLIPEPDRQRT
                        ******* *********** **
             Consensus  ASIILEFFLMVLLIPEPDRQRT
    Variable residues          L           L
```

B) Overlapping 9-mers generated from the three unique sequences represent all the inherent antigenic variations, with respect to potential 9-mer T-cell epitopes.

```
>37963458               >1854039                >17129648
ASIILEFLLMVLLIPEPDRQRT  ASIILEFFLMVLLIPEPDRQRT  ASIILEFFLMVLLIPEPDRLRT
asiilefll               ASIILEFFL               ASIILEFFL
 siilefllm               SIILEFFLM               SIILEFFLM
  iilefllmv               IILEFFLMV               IILEFFLMV
   ilefllmvl               ILEFFLMVL               ILEFFLMVL
    lefllmvll               LEFFLMVLL               LEFFLMVLL
     efllmvlli               EFFLMVLLI               EFFLMVLLI
      fllmvllip               FFLMVLLIP               FFLMVLLIP
       llmvllipe               FLMVLLIPE               FLMVLLIPE
        LMVLLIPEP               LMVLLIPEP               LMVLLIPEP
         MVLLIPEPD               MVLLIPEPD               MVLLIPEPD
          VLLIPEPDR               VLLIPEPDR               VLLIPEPDR
           LLIPEPDRQ               LLIPEPDRQ               llipepdrl
            LIPEPDRQR               LIPEPDRQR               lipepdrlr
             IPEPDRQRT               IPEPDRQRT               ipepdrlrt
```

Figure 3.1: Definition of antigenically redundant sequences. A) The three sequences (NCBI GI no.: 1854039, 17129648 and 37963458) are each unique, and residues that vary among them are shown. B) Overlapping 9-mers generated from the three unique sequences represent all the inherent antigenic variations, with respect to potential 9-mer T-cell epitopes. Although the three sequences are each unique, they share identical 9-mers. 9-mers shown in uppercase are those with an identical match in two of the unique sequences analyzed, while those in bold uppercase have an identical match in all three sequences; unique 9-mers are shown in lowercase. All the 9-mers in sequence 1854039 have a match in at least one of the other two sequences; thus, the antigenic diversity of this sequence can be covered by the other two sequences combined, rendering the sequence 1854039 antigenically redundant. Hence, the minimal number of sequences required to represent antigenic diversity for this dataset is two.

Table 3.4: Reduction of the number of unique dengue sequences by removal of antigenically redundant sequences.

Dengue serotype	Data retrieved in 2004			Data retrieved in 2005		
	Unique sequences (#)	Minimal antigenic set		Unique sequences (#)	Minimal antigenic set	
		Unique sequences (#) [a]	Percentage of unique sequences (%) [b]		Unique sequences (#) [a]	Percentage of unique sequences (%) [b]
DV1	359	244	68%	724	493	68%
DV2	507	368	73%	697	466	67%
DV3	230	180	78%	678	482	71%
DV4	222	177	80%	320	243	76%
Total	*1318*	*969*	*74%*	*2419*	*1684*	*70%*

[a] *Minimal number of unique sequences that represent complete short-peptide (9-mer) antigenic diversity of dengue unique sequences collected from the NCBI Entrez Protein database*

[b] *Percentage of unique sequences that represent complete short-peptide (9-mer) antigenic diversity of dengue unique sequences collected from the NCBI Entrez Protein database*

3.3.4 Characterization and application of sequence variables that affect antigenic diversity

The author examined the effects of sequence determinants, such as number and length of sequences, on short-peptide antigenic diversity of DV. These analyses were carried out using test datasets of different numbers of sequences (20, 40, 60, 80, 100, 120 and 140 sequences) and different lengths (23, 46, 128, 138, 276 and 460 aa) that were randomly selected from the set of DV2 envelope protein sequences, with repeated sampling of 20 times. Antigenic diversity analysis of each test dataset was performed to identify a minimal set of sequences that represents the complete short-peptide antigenic diversity for each dataset. These minimal sets were used to analyze the effects of the sequence determinants on antigenic diversity.

3.3.5 Effects of number of sequences on short-peptide antigenic diversity

An increase in the number of unique sequences in a dataset reduces the fraction required to represent the complete short-peptide antigenic diversity (Table 3.5). This observation reflects an asymptotic relationship between the number of unique sequences and the percentage of the complete short-peptide antigenic diversity that is covered (Figure 3.2). Asymptotic curves were observed for all the proteins of the four dengue serotypes (data not shown). The shape of the curve indicates that a single sequence will cover only a small proportion of the total short-peptide antigenic diversity and that for proteins with a large number of unique sequences, the addition of a single new variant sequence has little effect on the overall antigenic diversity.

Table 3.5: Effects of number of unique DV serotype 2 (DV2) envelope sequences (N) on short-peptide (9-mer) antigenic diversity. The mean and standard error (SE) values are shown for random repeated sampling of 20 times.

Number of unique sequences (N)	20	40	60	80	100	120	140
Length of sequences	460 aa	460 aa	460 aa	460 aa	460 aa	460 aa	460 aa
Minimal number of unique sequences that represent complete short-peptide antigenic diversity (Mean ± SE)	18 ± 0.30	32 ± 0.54	46 ± 0.70	58 ± 0.87	70 ± 0.87	80 ± 0.87	90 ± 0.71
Percentage of unique sequences that represent complete short-peptide antigenic diversity (%) (Mean ± SE)	90 ± 1.5	80 ± 1.35	77 ± 1.17	73 ± 1.09	70 ± 0.87	67 ± 0.73	64 ± 0.51

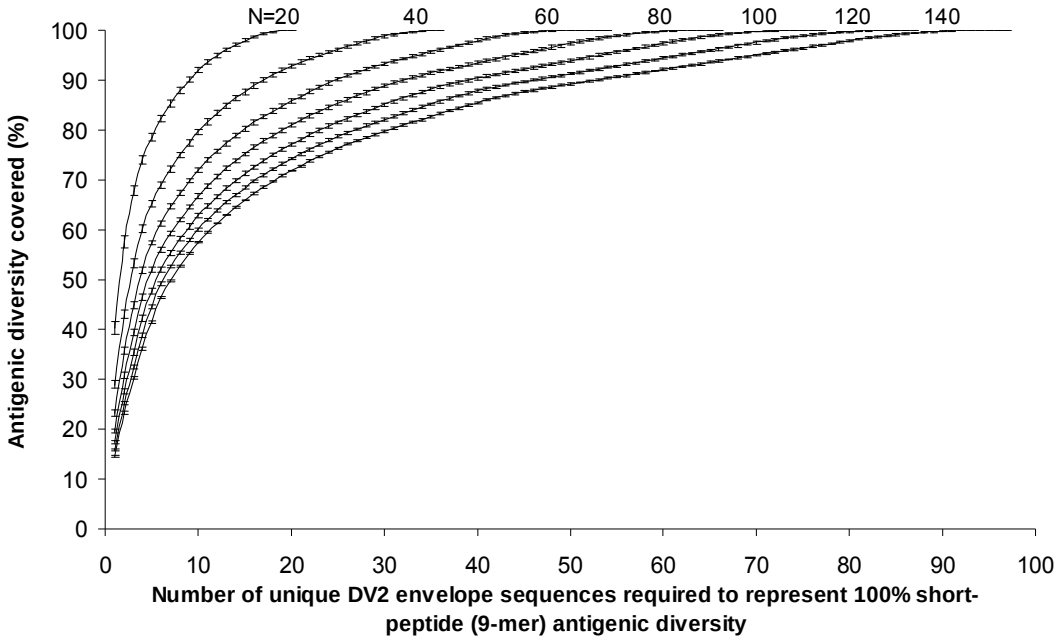


Figure 3.2: Short-peptide (9-mer) antigenic diversity as a function of number of sequences. Short-peptide antigenic diversity has an asymptotic relationship to number of unique DV serotype 2 (DV2) envelope sequences (N). Each curve shows the cumulative percentage coverage of short-peptide antigenic diversity. Vertical bars represent standard error for repeated random sampling of 20 times.

3.3.6 Effects of length of sequences on short-peptide antigenic diversity

A decrease in the length of sequences of a dataset reduces the fraction required to represent the complete short-peptide antigenic diversity of the dataset (Table 3.6). This reduction was achieved by removal of two types of redundancy: identical fragments and antigenically redundant fragments. The number of identical fragments increases significantly with a decrease in the length of the fragments because of the limited variability associated with smaller size. Hence, the effect of sequence length is significant, especially for very short fragments (23 aa), for which only ~7% of the unique fragments were required to represent complete antigenic diversity of the short fragments (a reduction of ~93%). Overall, the results indicate that short-peptide antigenic diversity has a near-linear relationship to sequence length (Figure 3.3).

Table 3.6: Effects of length of DV serotype 2 (DV2) envelope protein sequences on short-peptide (9-mer) antigenic diversity. The mean and standard error (SE) values are shown for random repeated sampling of 20 times.

Length of fragments	100% (460 aa)	60% (276 aa)	30% (138 aa)	20% (92 aa)	10% (46 aa)	5% (23 aa)
Number of fragments	187	187	187	187	187	187
Number of unique fragments	187	131	82	58	27	17
Minimal number of fragments that represent complete short-peptide antigenic diversity (Mean ± SE)	111 ± 0.11	74 ± 0.11	48 ± 0.17	38 ± 0.10	24 ± 0.10	14 ± 0.10
Percentage of fragments that represent complete short-peptide antigenic diversity (%) (Mean ± SE)	59 ± 0.06	40 ± 0.06	26 ± 0.09	20 ± 0.05	13 ± 0.05	7 ± 0.05

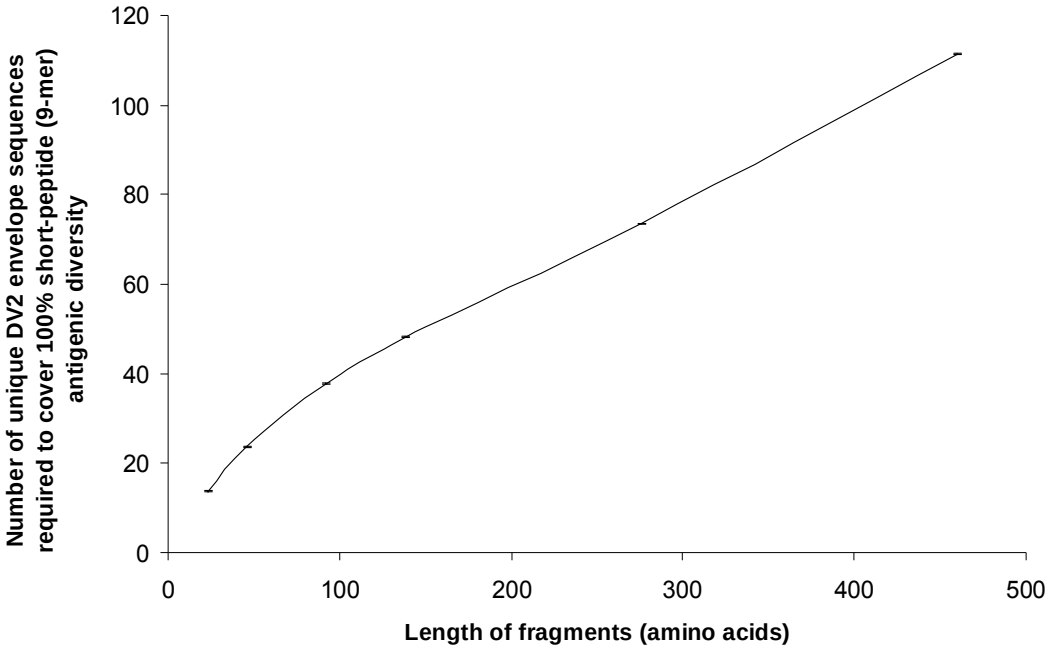


Figure 3.3: Short-peptide (9-mer) antigenic diversity as a function of length of sequences. Short-peptide antigenic diversity shows a linear relationship to the sequence length of DV serotype 2 (DV2) envelope protein.

3.3.7 Summary of results

In summary, the number of unique protein sequences required to represent complete antigenic diversity of short peptides in DV was significantly smaller than that required

to represent complete protein sequence diversity (Figure 3.4). Assessment of factors that determine antigenic diversity revealed asymptotic relationship of short-peptide antigenic diversity to number of unique protein sequences and near-linear relationship to sequence length.

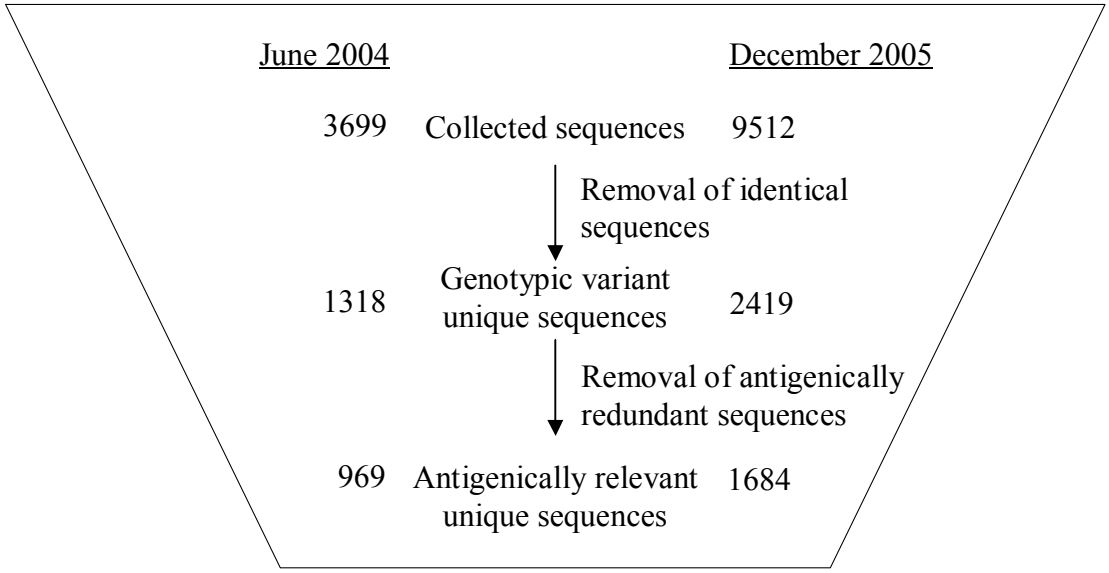


Figure 3.4: Flowchart summarizing the steps undertaken to identify the antigenically relevant unique sequences in the DV.

3.4 Discussion

In this study, a systematic bioinformatics approach was applied to collect, clean, organize and analyze the antigenic diversity of short peptides in reported protein sequence data of DV. A computational method was developed for the analysis of antigenic diversity of short peptides within DENV proteins. This method was applied for the analysis of short-peptide antigenic diversity of DV to determine a minimal sequence set that encodes the complete antigenic diversity of linear epitopes within each DV serotype. The relationship between short-peptide antigenic diversity and protein sequence diversity of DV were studied and the effects of sequence

determinants on viral antigenic diversity were also explored. Our analysis showed that the minimal number of unique sequences required to represent complete antigenic diversity of linear epitopes in DV is significantly smaller than that required to represent complete protein sequence diversity. Short-peptide antigenic diversity shows an asymptotic relationship to the number of unique sequences and linear relationship to the length of protein antigens.

The minimal sequence set that encodes the complete short-peptide antigenic diversity for each DV serotype was derived through removal of identical sequences and antigenically redundant sequences (Table 3.4 and Figure 3.4). Both reductions occurred without any loss of information on antigenic diversity among the sequences. The largest reduction was accomplished through the removal of identical sequences, since only 36% (year 2004) or 25% (year 2005) of the sequences were unique. The identical sequences originated from DV strains that were unique variants with respect to the whole polyprotein, but were identical to other dengue strains with respect to individual proteins, resulting in many duplicate protein sequences. The removal of antigenically redundant sequences also involved a significant proportion of the sequences, approximately one-third of all unique sequences (2004: 26%; 2005: 30%), reflecting the high antigenic redundancy among the DV variants, which often differed by only a few amino acids. Despite significant reduction achieved by reducing the collected sequences to minimal sequences, a large number of protein sequences, 969 in 2004 and 1684 in 2005, were still required to represent the complete short-peptide antigenic diversity of DV.

It is clear that antigenic diversity in the reported dengue sequences is large. With many asymptomatic human and animal carriers of DVs representing a huge reservoir for emergence of new strains (Schein *et al.*, 2005; Holmes and Twiddy,

2003; Halstead and Deen, 2002), the diversity is expected to increase, although at a progressively slower pace. This is because antigenic redundancy increases when the number of sequences increases; it was observed that when the dataset for a particular protein reaches approximately 200 sequences, the effect of addition of new sequences to increasing antigenic diversity is marginal.

Our study of factors that affect antigenic diversity provided insight into dealing with the increasing T-cell epitope antigenic diversity in the context of vaccine development. Length of sequences had the largest effect on short-peptide antigenic diversity. Asymptotic behaviour of antigenic diversity increase was observed for the increase in the number of sequence variants. Selection of the region of the sequence also had some effect (data not shown) – the higher the conservation of the region, the lesser the diversity, with complete conservation being ideal. For practical purposes of vaccine formulation, antigenic diversity cannot be represented by whole protein sequences because it is not feasible to use these sequences for systematic experimental analysis: they are long and their number is increasing rapidly. The implication is that conventional vaccination strategies, which utilize whole attenuated pathogen with little knowledge of the specificity of immune responses they elicit, may not be suitable for providing protection from multiple variants of viruses. Furthermore, it may be difficult to optimize such vaccine according to the HLA profile of the population receiving the vaccine (Brusic and August, 2004; Ovsyannikova *et al.*, 2004), as neither the T-cell epitopes and nor their HLA restrictions are known.

A more effective vaccine strategy that the author proposes is to focus on short, conserved segments of proteins (~ <100 aa) that are known to be specific targets of immune responses (such as T-cell epitopes specific to particular HLA alleles). To deal

with the diversity of the immune system, it is important to map the T-cell epitopes relative to the HLA supertypes (Sette and Sidney, 1999; Sette *et al*., 2001), such as T-cell epitopes promiscuous to HLA-DR or other supertypes. For a set of sequences, by identifying the minimal number of short, conserved peptides that represent antigenic diversity relevant to each supertype and combining them, the complete antigenic diversity relevant to multiple supertypes can be covered in a "divide-and-conquer" approach. This may provide a promising basis for multivalent peptide-based vaccine against DV. The concept of using conserved, supertype restricted epitopes to target pathogen antigenic diversity and as peptide-based vaccine targets is also supported by others (Sette *et al*., 2001; Sylvester-Hvid *et al*., 2002; De Groot *et al.*, 2005).

There are some caveats to be considered in this study. First, it is well-known that not all HLA-restricted epitopes are 9-mers (Rammensee, 1995). This may impact the interpretation of our results, which were based only on 9-mers, and hence may not give a true representation of dengue T-cell epitope antigenic diversity. 9-mers were selected because they represent the typical size of HLA class I T-cell epitopes, as well as the binding core of HLA class II T-cell epitopes (Rammensee, 1995). Nevertheless, similar analysis with peptides of 8-mers and 10-mers showed no significant difference as compared to the analysis of 9-mers (data not shown).

The second caveat is the sampling bias in DV sequences reported in the public databases. Only dengue sequences that have been studied are reported, and viruses collected in accessible locations, associated with notable disease outbreaks or of known immunological properties are preferentially studied. Consequently, certain dengue proteins have been studied intensively, while the others remained largely unstudied. For example, sequences of the envelope protein, known to be important for immunological activity and viral entry into host (Kurane *et al.*, 1998; Preugschat and

Strauss, 1991), were the most abundant in our dataset (3183 sequences for all four serotypes), while that of NS4a, which is relatively unknown for immunological activity, was under-represented. In addition, for majority of the proteins, a large portion of the reported sequences were incomplete in length. For example, 95% of DV2 NS5 collected sequences were incomplete in length (data not shown). However, the data used in this study was the most representative available and the large sample size for majority of the proteins helps to decrease the margin of error due to sampling bias. In addition, the reported sequences represent highly pathogenic strains isolated during dengue outbreaks.

3.5 Conclusions

This study has provided evidence that there are limited number of antigenic combinations in variant protein sequences of a viral species and that short regions of the viral proteins are sufficient to cover antigenic diversity of T-cell epitopes. The approach described here has direct application to the analysis of other viruses, in particular those that show high diversity and/or rapid evolution, such as influenza A virus and HIV. Preliminary results of analysis to other viruses by my colleagues in the lab (Mr. Kenneth Lee Xunjian on West Nile virus and Ms. Heiny Tan on Influenza A), applying the computational methodology developed, showed similar result of limited antigenic combination in variant protein sequence of a viral species.

3.6 Chapter summary

Background: Antigenic diversity in DV strains has been studied, but large-scale and detailed systematic analyses have not been reported. In this study, a bioinformatics

method for analyzing viral antigenic diversity in the context of T-cell mediated immune responses is described. The method was applied to study the relationship between short-peptide antigenic diversity and protein sequence diversity of DV. The effects of sequence determinants on viral antigenic diversity were also studied. Short peptides, principally 9-mers were studied because they represent the predominant length of binding cores of T-cell epitopes, which are important for formulation of vaccines.

***Results**:* The analysis showed that the number of unique protein sequences required to represent complete antigenic diversity of short peptides in DV was significantly smaller than that required to represent complete protein sequence diversity. Short-peptide antigenic diversity showed an asymptotic relationship to the number of unique protein sequences, indicating that for large sequence sets (~200) the addition of new protein sequences has marginal effect to increasing antigenic diversity. A near-linear relationship was observed between the extent of antigenic diversity and the length of protein sequences, suggesting that, for the practical purpose of vaccine development, antigenic diversity of short peptides from DV can be represented by short, conserved regions of sequences (~<100 aa) within viral antigens that are specific targets of immune responses (such as T-cell epitopes specific to particular HLA alleles), in particular promiscuous T-cell epitopes.

***Conclusions**:* This study provides evidence that there are limited number of antigenic combinations in protein sequence variants of a viral species and that short, conserved regions of the viral protein are sufficient to cover antigenic diversity of T-cell epitopes. The approach described herein has direct application to the analysis of other viruses, in particular those that show high diversity and/or rapid evolution, such as influenza A virus and human immunodeficiency virus (HIV).

Chapter 4 Identification and Characterization of Dengue Virus Peptides that Cover Antigenic Diversity (*PE*s)

4.1 Introduction

Based on the insights gained from the analysis presented in Chapter 3, we have focused on short, conserved sequence fragments that contain promiscuous T-cell epitopes to cover antigenic diversity, for the practical purpose of vaccine development. Therefore, the author analysed all available DENV sequence data in public databases to identify and characterize peptides that cover antigenic diversity (*PE*s) of the virus: sequence regions conserved across sequences of the four DENV serotypes (pan-DENV sequences) and are immunologically relevant in the context of HLA supertypes.

Bioinformatics-based approaches were used to (a) extract all DENV sequences available in public databases (as of December 2007), (b) identify pan-DENV sequences, (c) analyze the evolutionary stability of the pan-DENV sequences, (d) characterize the structure-function relationship and distribution in nature of the pan-DENV sequences, and (e) examine the immune relevance of the conserved sequences as potential promiscuous T-cell epitopes that are applicable to the majority of the human population worldwide (Sette and Sidney, 1999). The pan-DENV sequences were also correlated to previously reported T-cell epitopes and those identified in HLA transgenic mice (Tg) by a collaborator of the author.

The author clarifies here that the bioinformatics analyses, which are the scope of this thesis, were performed by the author himself and that the experimental validation was done independently as a full-overlapping study, the results of which the author used to validate his findings.

4.2 Materials and methods

4.2.1 Methodology overview

The bioinformatics approaches adopted in this study are summarized in Figure 4.1. These include three major steps that involve creating the database, analysis of potential T-cell epitopes, and experimental validation.

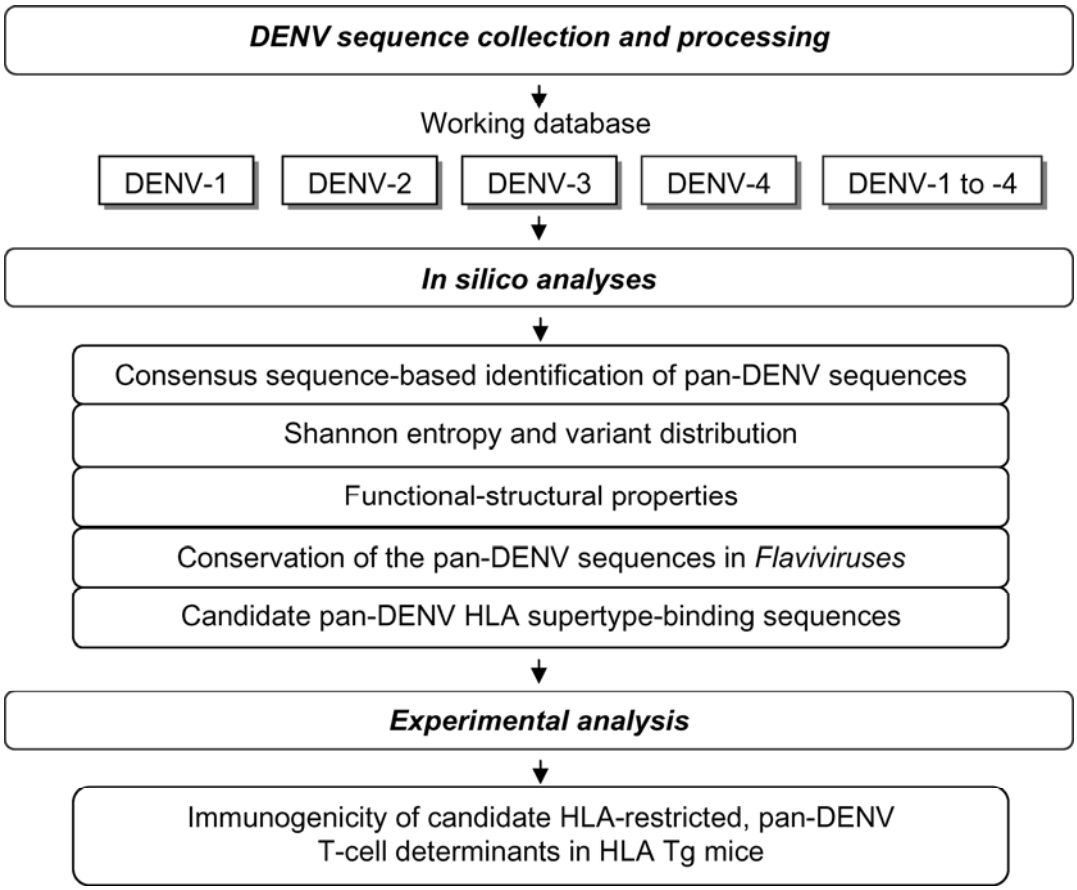


Figure 4.1: Overview of bioinformatics and experimental approaches employed for identification and analysis of pan-DENV sequences.

4.2.2 Dengue virus data collection and sequence organization

DENV protein sequences were retrieved from the NCBI Entrez Protein database in December 2005 (included reported data as of December 2005), and again in December 2007 (included reported data as of December 2007) for validation purposes, by use of a taxonomy ID search via the NCBI taxonomy browser (Wheeler *et al.*, 2005). The taxonomy IDs for DENV-1 to -4 are 11053, 11060, 11069 and 11070, respectively. The data for 2007 were processed separately from the 2005 dataset, but using identical procedures.

The sequences for the proteins C, prM, E, NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5 of each serotype were extracted from the database records and grouped according to the steps described in Chapter 3.2.2. The resulting datasets for each serotype protein were then aligned by use of ClustalX 1.83 (Thompson *et al.*, 1997) with default parameters, followed by manual inspection and correction of misalignments. The alignments, which comprised both full length and partial sequences, were then subjected to a number of analyses. Identical sequences were not removed from the alignments, unless otherwise indicated in the sections below, because they reflected the incidence of the corresponding DENV isolates in nature.

4.2.3 Identification of pan-DENV sequences

The DENV protein sequences were examined by a consensus-sequence based approach (Novitsky *et al.*, 2002) to identify sequence fragments that were common across the four serotypes. The consensus-sequence approach was used because it helps compress a multiple sequence alignment of a protein into a single consensus sequence encoding only predominant residues. This facilitates identification of

sequence fragments across the serotype proteins that are not only common but also highly represented in each serotype.

The consensus sequences for the proteins of each serotype (intra-serotype consensus) were first derived by multiple sequence alignments to select the predominant residue at each amino acid position. The four intra-serotype consensus sequences for a given protein (one from each serotype) were then aligned to reveal sequence fragments, at least nine amino acids long that were identical across each of the serotypes. This minimum length was chosen because it represents the binding core length of a majority of HLA-restricted T-cell epitopes (Rammensee *et al.*, 1995). Only sequence fragments that were identical in at least 80% of the sequences of each of the four serotypes were retained for further analyses. The 80% intra-serotype representation cut-off was chosen because 44 of the 46 sequence fragments that were common across the four DENV serotypes exhibited intra-serotype representation of ≥ 81%, and the two that did not, had significantly lower representation (~56-67%) in one of the four serotypes. Peptides with residue X in the alignment were ignored from the percentage representation (*i.e.* frequency) computation.

4.2.4 Entropy analysis of pan-DENV sequences

Shannon's entropy (Shannon, 1948) was used to quantify the diversity of DENV protein sequences within each serotype (intra-serotype diversity) and across all DENVs (pan-DENV diversity), and to assess the predicted evolutionary stability of the identified pan-DENV sequences. All entropy analyses were carried out by using the in-house developed Antigenic Variability Analyser tool (AVANA; for which the author of this thesis contributed significantly with user specifications and testing cases) (Miotto *et al.*, 2008). For immunological applications, the entropy measure for

antigenic sequences was based on nonamer peptides (Rammensee *et al.*, 1995), centered at any given position in the alignment. Applying Shannon's formula, the nonamer peptide entropy *H(x)* at any given position *x* in the alignment was computed by

$$H(x) = -\sum_{i=1}^{n(x)} p(i,x) \log_2 p(i,x)$$

where *p(i, x)* is the probability of a particular nonamer peptide *i* being centered at position *x*. The entropy value increases with *n(x)*, the total number of peptides observed at position *x*. It is also affected by the relative frequency or probability of the peptides, such that it decreases when one peptide is highly represented in the sequences. In theory, nonamer entropy values can range from 0, for a position with completely conserved nonamer peptide in all sequences analyzed, to 39 ($\log_2 20^9$); in practice, however, the upper bound is very much lower as natural sequences of a protein family tend to be closely related. Currently available highly diverse HIV protein sequences peak at an entropy value of eight (data not shown).

Since peptide entropy is computed at a nonamer's center position, the first and last four positions in each protein alignment are not assigned peptide entropy values. Only sequences that contain a valid amino acid at position *x* in an alignment are used for the entropy computation, and nonamers containing only gaps are ignored. Although gaps tend to occur in high-diversity regions, proteins that have a high fraction of gaps have reduced statistical support, yielding artificially low entropy values; thus, positions with more than 50% of the sequences containing gaps are discarded. Both complete and partial protein sequences can be used in the entropy computation because of the statistical nature of the entropy measure.

For finite-size sets of sequences, entropy computations are affected by the sequence count in the alignment. For an alignment of N sequences, alignment size

bias is proportional to 1/N (Paninski, 2003). This relationship allows a correction for size bias by applying to each alignment a statistical adjustment that estimates entropy values for an infinitely-sized alignment with analogous peptide distribution. To obtain such an estimate, the alignment was repeatedly randomly sampled to create smaller alignments of varying size, whose entropy was measured. At each alignment position, the entropy of these subset alignments of size N was plotted against 1/N, using a linear regression to extrapolate the entropy estimate for $N \rightarrow \infty$. The regression's coefficient of determination (r^2) was used as a goodness-of-fit of the resulting estimate. In this study, size bias correction was applied to all entropy calculations, so that alignment sequence counts could be ignored in comparisons. All entropy values reported herein are therefore infinite-size set estimates, rather than the values directly computed from the alignments.

4.2.5 Nonamer variant analysis of pan-DENV sequences

Data from entropy analysis were used to study the distribution of the representation of nonamer variant peptides in pan-DENV and non pan-DENV sequences regions, within and across the serotypes. Variant nonamers for a given position *x* in the alignment were defined as all nonamers that differed by at least one amino acid from the predominant nonamer (peptide that was contained in the majority of the sequences) at the position. Therefore, for any given position *x* in the alignment, the combined representation of all nonamers was computed by subtracting the percentage representation of the predominant peptide from 100%.

4.2.6 Functional and structural analyses of pan-DENV sequences

The known and putative structural and functional properties of pan-DENV sequences were searched in the literature and by use of the Prosite (Hulo *et al.*, 2006), via ScanProsite (de Castro *et al.*, 2006), and Pfam (Bateman *et al.*, 2004) databases. When possible, the sequences were mapped on the three-dimensional (3-D) structures of available DENV antigen (Ag) in the Protein Data Bank (PDB) (Berman *et al.*, 2000) (www.pdb.org) by use of ICM-Browser version 3.3 (www.molsoft.com). X-ray diffraction 3-D structures were visualized by use of the Corey, Pauling and Koltun (cpk) representation in the ICM-Browser.

4.2.7 Identification of pan-DENV sequences common to other viruses and organisms

Pan-DENV sequences that overlapped at least nine consecutive amino acid sequences of other viruses and organisms were identified by performing BLAST search against all viral protein sequences reported at NCBI (as of July 2007), excluding DENV sequences (parameters set: limit by Entrez query "txid10239[Organism:exp] NOT txid12637[Organism:exp]"; checked the option "automatically adjust parameters for short sequences"; unchecked the "low-complexity filter"; alignment option set to a maximum of "20,000" hits). Similar BLAST searches were carried out against protein sequences of all organisms excluding viruses (parameters set: limit by Entrez query "Root[ORGN] NOT Viruses[ORGN] NOT txid81077[ORGN]"; checked the option "automatically adjust parameters for short sequences"; unchecked the "low-complexity filter"; alignment option set to a maximum of "20,000" hits). The keyword "NOT txid81077 [ORGN]" was used to remove artificial sequence hits.

4.2.8 Identification of known and predicted pan-DENV HLA supertype binding sequences

Both literature search and query against the Immune Epitope Database (Peters *et al.*, 2005) (www.immuneepitope.org) were performed to detect reported immunogenic, human T-cell epitopes (both class I and II) of DENV that either fully or partially overlapped with the pan-DENV sequences. In addition, dedicated algorithms based on several prediction models were used to identify candidate putative HLA-binding sequences to multiple HLA class I and II supertype alleles within the pan-DENV sequences. Putative HLA supertypes class I-restricted peptides were identified by use of NetCTL (Larsen *et al.*, 2005), Multipred (Zhang *et al.*, 2005b), ARB (Bui *et al.*, 2005), and class II-restricted peptides by Multipred and TEPITOPE (Bian and Hammer, 2004). Further, the intra-serotype representation of the putative T-cell epitopes was analyzed.

The NetCTL 1.2 algorithm (www.cbs.dtu.dk/services/NetCTL) predicts peptides restricted by 12 HLA class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 and B62). The algorithm integrates the predictions of HLA binding, proteasomal C-terminal cleavage and transport efficiency by the transporter associated with antigen processing (TAP) molecules. HLA binding and proteasomal cleavage predictions are performed by an artificial neural networks (ANN) method, while TAP transport efficiency is predicted using a weight matrix method. The parameters used for NetCTL prediction were: 0.15 weight on C terminal cleavage (default), 0.05 weight on TAP transport efficiency (default), and 0.5 threshold for HLA supertype binding, which was reported to be optimal (sensitivity (SN), 0.89 and specificity (SP), 0.94) in a large benchmark study containing more than 800 known class I T-cell epitopes (Larsen *et al.*, 2005).

The TEPITOPE software (2000 beta version; courtesy of J. Hammer) utilizes quantitative matrix-based motifs, obtained from experimental scanning of the binding of P1-anchored designer peptides to soluble HLA-DR molecules in *in-vitro* competition assays, to predict peptides binding to 25 common HLA-DR alleles (DRB1*0101, *0102, *0301, *0401, *0402, *0404, *0405, *0410, *0421, *0701, *0801, *0802, *0804, *0806, *1101, *1104, *1106, *1107, *1305, *1307, *1311, *1321, *1501, *1502, and DRB5*0101) (Bian and Hammer, 2004; Sturniolo *et al.*, 1999). The parameters for TEPITOPE predictions were: 5% quantitative threshold and putative epitopes with a 10-fold inhibitory residue included. Nonamer peptides predicted to bind at least 10 out of the 25 HLA-DR alleles were selected as putative supertype-restricted epitopes.

Multipred is a computational system for the prediction of peptides that bind to HLA class I supertypes A2 and A3 and class II HLA-DR supertype (Zhang *et al.*, 2005b) (the author was significantly involved in the development of this system). The HLA alleles selected to represent these supertypes by Multipred were as follows: A2 supertype, A*0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209; A3 supertype, A*0301, *0302, *1101, *1102, *3101, *3301 and *6801; DR supertype, DRB1*0101, *0301, *0401, *0701, *0801, *1101, *1301, and *1501. Hidden Markov model (HMM) and ANN methods are the predictive models of Multipred; both have been optimized and show similar performances (Zhang *et al.*, 2005b). The sum thresholds used for prediction of peptides restricted to the three HLA supertypes by ANN and HMM methods were: A2, 31.33 (ANN; SN = 0.80 and SP = 0.83) and 47.08 (HMM; SN = 0.80 and SP = 0.78); A3, 24.53 (ANN; SN = 0.90 and SP = 0.95) and 37.58 (HMM; SN = 0.80 and SP = 0.87); and DR, 23.42 (ANN; SN = 0.90 and

SP = 0.92) and 51.08 (HMM; SN = 0.90 and SP = 1.00). Consensus predictions of the two methods were taken as final predictions for each HLA supertype.

The ARB matrix (epitope.liai.org:8080/matrix/matrix_prediction.jsp) method is based on a matrix of coefficients to predict IC50 values (Bui *et al.*, 2005). The HLA class I alleles predicted by ARB were grouped according to the current supertype classification (Sette *et al.*, 2003; Sette and Sidney, 1999) and supertypes containing more than two alleles predicted by the system were selected, namely A2 (A*0201, *0202, *0203, *0206, and *6802), A3 (A*0301, *1101, *3101, *3301 and *6801), B7 (B*0702, A*3501, *5101, *5301, and *5401), and B44 supertypes (B*4001, *4002, *4402, *4403, and *4501). The prediction threshold value chosen for optimum sensitivity and specificity was IC50 $\leq$ 1000 nM and nonamer peptides predicted to bind three or more alleles of the supertype were considered as putative promiscuous HLA supertype-restricted epitopes.

4.2.9 ELISpot analysis of HLA-DR restricted epitopes in pan-DENV sequences

Experimental validation was performed to examine whether the pan-DENV sequence contained targets of cellular immune responses. All the experiments were kindly performed by our collaborator Dr. Eduardo Nascimento from the Johns Hopkins University School of Medicine, in Baltimore, Maryland, USA. All experiments were approved by the Johns Hopkins University Institutional Animal Care and Use Committee. Murine H-2 class II-deficient, HLA-DR2 (Vandenbark *et al.*, 2003), HLA-DR3 (Madsen *et al.*, 1999; Strauss *et al.*, 1994), HLA-DR4 (referred to as DR4/IE) (Ito *et al.*, 1996) and HLA-DR4/human CD4 (huCD4) (Cope *et al.*, 1999; Fugger *et al.*, 1994) transgenic mice were used, bred and maintained in the Johns

Hopkins University School of Medicine Animal Facility. Specific pathogen-free (SFP) colonies were maintained in a helicobacter-negative mice facility. The HLA-DR expression of the experimental transgenic mice was evaluated by flow cytometry.

Mice were immunized subcutaneously at the base of the tail, twice at two weeks interval, with pools of overlapping peptides covering the DENV-3 protein (15-17 aa, overlapping by 10-11 aa) (Schafer-N Inc., Copenhagen, Denmark; BEI Resources, Manassas, VA). Peptide pools (73-155 peptides per pool) contained 1 μg of each peptide and were emulsified (1:1) in TiterMax adjuvant (TiterMax USA, Inc.). An aqueous preparation of TiterMax (1:1) was used as a negative control. Two weeks after the second immunization, the mice were sacrificed and HLA-DR-restricted CD4 T-cell responses were assessed by *ex vivo* IFN-γ ELISpot assay using CD8-depleted splenocytes. Each target peptide was tested in duplicate. Spot-forming cell (SFC) counts were normalized to 10^6 cells. The results were considered significant when the average SFC minus two standard deviations (SD) was greater than the average of the background plus 2 SD; and the average values were greater than 10 SFC per 10^6 splenocytes. The initial screening assays were performed with peptide matrices (Roederer and Koup, 2003), followed by assays with the relevant individual peptides (Nascimento *et al.*, manuscript in preparation).

4.3 Results

4.3.1 Dengue virus serotype protein datasets

A total of 9,512 and 12,404 complete and partial DENV protein sequences were collected from the NCBI Entrez Protein database in December 2005 and again in December 2007, respectively, representing an increase of approximately 30% (2892

sequences) in the 24-months interval (Table 4.1). The total number of sequences (2007) varied from 4,011 for DENV-2 to 1,415 for DENV-4 and from 3,845 for E to 523 for NS4a proteins. Most of the individual protein sequences originated from DENV strains that were unique variants with respect to the entire polyprotein, but were identical to other strains with respect to individual proteins (Khan *et al.*, 2006a).

4.3.2 Conserved pan-DENV sequences

The consensus-sequence approach identified a total of 44 pan-DENV sequences of at least nine amino acids that were present in ≥ 80% of all sequences of each DENV serotype for both 2005 and 2007 datasets (Figure 4.2; Table 4.2). Strikingly, 34 of the 44 (~77%) were conserved in ≥ 95% of all reported DENV sequences. The size of the pan-DENV sequences ranged from nine to 22 amino acids, with a combined size of 514 residues, corresponding approximately to 15% of the complete DENV polyprotein (~3390 amino acids) (Table 4.3). The vast majority (42/44) of the pan-DENV sequences were localized in the NS proteins, with 17, 12, 7 and 5 sequences found in NS5, NS3, NS1 and NS4b, respectively, and 1 in the NS4a protein. Notably, the remaining two pan-DENV sequences were localized in the E protein. No pan-DENV sequence was found in the C, prM, NS2a and NS2b proteins. The largest size of the combined pan-DENV sequences was in the NS5 protein, representing a total of 215 amino acid positions covering ~24% of the protein, followed by NS3, NS1 and NS4b with 122, 74 and 69 amino acid positions covering ~20, ~21 and ~28% of the corresponding proteins, respectively. The two pan-DENV sequences in the E protein had a combined size of only 25 amino acids, corresponding to ~5% of the protein.

Table 4.1: Number and distribution of reported DENV protein sequences.

DENV protein [b]	No. of sequences [a]										
	DENV-1		DENV-2		DENV-3		DENV-4		*Total*		
	2005	2007	2005	2007	2005	2007	2005	2007	2005	2007	Increase
C	194	298	266	311	414	547	117	122	991	1278	287
prM	206	311	353	404	458	590	207	225	1224	1530	306
E	852	1051	1277	1518	716	910	338	366	3183	3845	662
NS1	410	565	640	752	201	308	142	159	1393	1784	391
NS2a	150	238	132	173	90	169	121	125	493	705	212
NS2b	136	224	130	163	104	183	40	44	410	614	204
NS3	98	186	145	178	216	297	30	34	489	695	206
NS4a	91	178	128	162	70	151	28	32	317	523	206
NS4b	89	176	129	163	70	150	109	113	397	602	205
NS5	92	179	151	187	181	267	191	195	615	828	213
Total	2318	3406	3351	4011	2520	3572	1323	1415	9512	12404	2892

[a] *Collected from the NCBI Entrez Protein database*

[b] *Manually processed after multiple sequence alignments and use of the known DENV cleavage sites*


DENV-1
DENV-3
DENV-2
DENV-4
Percentage representation (%)
80
90
100
E
97-VDRGWGNGCGLFGKG-111
252-VLGSQEGAMH-261
NS1
12-ELKCGSGIF-20
25-VHTWTEQYKFQ-35
193-AVHADMGYWIES-204
229-HTLWSNGVLES-239
266-GPWHLGKLE-274
294-RGPSLRTTT-302
325-GEDGCWYGMEIRP-337
NS3
46-FHTMWHVTRG-55
148-GLYGNGVVT-156
189-LTIMDLHPG-197
256-EIVDLMCHATFT-267
284-MDEAHFTDP-292
296-AARGYISTRV-305
313-IFMTATPPG-321
357-GKTVWFVPSIK-367
383-VIQLSRKTFD-392
406-VVTTDISEMGANF-418
491-EAKMLLDNI-499
537-LMRRGDLPVWL-547
NS4a
126-QRTPQDNQL-134
NS4b
35-PASAWTLYAVATT-47
118-HYAIIGPGLQAKATREAQKR-137
139-AAGIMKNPTVDGI-151
213-FWNTTIAVS-221
223-ANIFRGSYLAGAGL-236
NS5
6-GETLGEKWK-14
79-DLGCGRGGWSYY-90
104-TKGGPGHEEP-113
141-DTLLCDIGESS-151
209-PLSRNSTHEMYW-220
302-TWAYHGSYE-310
342-AMTDTTPFGQQRVFKEKVDTRT-363
450-CVYNMMGKREKKLGEFG-466
468-AKGSRAIWYMWLGAR-482
505-SGVEGEGLH-513
531-YADDTAGWDTRIT-543
568-IFKLTYQNKVV-578
597-DQRGSGQVGTYGLNTFTNME-616
658-RMAISGDDCVVKP-670
707-VPFCSHHFH-715
765-LMYFHRRDLRLA-776
790-PTSRTTWSIHA-800

Figure 4.2: Pan-DENV sequences and their representations in the four DENV serotypes. The 44 pan-DENV sequences of at least nine amino acids that were found present in ≥ 80% of the recorded sequences of each DENV serotype are shown. The representation values are shown for the 2005 dataset; see Table 4.2 for values of both 2005 and 2007 datasets. Amino acid positions were numbered according to the sequence alignments of the four DENV serotypes. The corresponding proteins are indicated on the left.

In large-scale proteome analyses such as this study, bias may result from the collection of completely or partially overlapping redundant sequences, corresponding to identical or highly similar circulating DENV isolates sequenced by various dengue surveillance programs in different countries. Although to some extent this redundancy may be accepted as a reflection of the incidence of the corresponding DENV isolates in nature, the author assessed its potential bias effect by repeating the analysis of conservation after discarding duplicate sequences from the datasets. The analysis of unique sequences identified all the same pan-DENV sequences that were identified when including duplicates (Figure 4.2), except for $NS1_{12-20}$, $NS1_{25-35}$ and $NS5_{597-616}$. Therefore, the presence of duplicates in the DENV datasets did not significantly affect the results. Although the removal of duplicates does not fully compensate for biases in the datasets, the removal of highly similar sequences, which may have been generated from relatively large sequencing efforts in single outbreaks, was deemed undesirable, since such arbitrary selection would introduce additional bias.

Table 4.2: The intra-serotype percentage representation of pan-DENV sequences.

DENV protein	Pan-DENV sequence [a]	% intra-serotype representation [b]							
		DENV-1		DENV-2		DENV-3		DENV-4	
		2005	2007	2005	2007	2005	2007	2005	2007
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$	97.8	98.2	98.0	98.3	100.0	99.8	95.0	95.4
	$_{252}$VLGSQEGAMH$_{261}$	95.5	96.4	98.5	98.3	98.1	98.6	99.7	99.7
NS1	$_{12}$ELKCGSGIF$_{20}$	99.5	99.2	94.9	94.3	83.1	85.9	98.6	98.7
	$_{25}$VHTWTEQYKFQ$_{35}$	98.6	99.0	95.1	95.3	99.0	98.6	95.0	94.8
	$_{193}$AVHADMGYWIES$_{204}$	100.0	99.5	94.9	95.5	96.5	93.6	96.4	93.8
	$_{229}$HTLWSNGVLES$_{239}$	96.8	97.7	95.8	96.3	96.6	98.1	100.0	97.0
	$_{266}$GPWHLGKLE$_{274}$	100.0	100.0	94.4	92.1	99.1	99.5	100.0	100.0
	$_{294}$RGPSLRTTT$_{302}$	93.7	95.9	99.1	99.3	98.3	98.5	100.0	100.0
	$_{325}$GEDGCWYGMEIRP$_{337}$	98.1	98.0	97.0	97.1	98.2	99.0	100.0	100.0
NS3	$_{46}$FHTMWHVTRG$_{55}$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	$_{148}$GLYGNGVVT$_{156}$	100.0	100.0	99.3	99.4	100.0	100.0	100.0	100.0
	$_{189}$LTIMDLHPG$_{197}$	100.0	100.0	98.6	98.9	100.0	100.0	100.0	97.0
	$_{256}$EIVDLMCHATFT$_{267}$	99.0	99.5	100.0	100.0	100.0	100.0	100.0	100.0
	$_{284}$MDEAHFTDP$_{292}$	98.9	98.9	100.0	100.0	100.0	100.0	100.0	100.0
	$_{296}$AARGYISTRV$_{305}$	96.7	97.7	100.0	100.0	97.3	98.1	100.0	100.0
	$_{313}$IFMTATPPG$_{321}$	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0
	$_{357}$GKTVWFVPSIK$_{367}$	98.9	99.4	100.0	91.7	99.4	99.6	96.3	96.8
	$_{383}$VIQLSRKTFD$_{392}$	81.1	89.8	98.5	98.8	98.3	98.9	100.0	100.0
	$_{406}$VVTTDISEMGANF$_{418}$	97.8	98.9	98.5	98.8	97.8	98.5	100.0	100.0
	$_{491}$EAKMLLDNI$_{499}$	96.7	98.3	100.0	100.0	99.4	99.6	100.0	100.0
	$_{537}$LMRRGDLPVWL$_{547}$	98.9	99.4	100.0	92.2	99.4	99.2	92.6	90.3
NS4a	$_{126}$QRTPQDNQL$_{134}$	97.7	98.9	100.0	100.0	100.0	100.0	100.0	100.0
NS4b	$_{35}$PASAWTLYAVATT$_{47}$	100.0	100.0	100.0	100.0	100.0	99.3	100.0	100.0
	$_{118}$HYAIIGPGLQAKATREAQKR$_{137}$	98.9	98.9	95.3	95.7	100.0	100.0	98.2	98.2
	$_{139}$AAGIMKNPTVDGI$_{151}$	95.5	97.7	97.6	97.5	97.1	98.7	100.0	100.0
	$_{213}$FWNTTIAVS$_{221}$	97.7	98.9	100.0	100.0	100.0	100.0	98.2	98.2
	$_{223}$ANIFRGSYLAGAGL$_{236}$	100.0	100.0	100.0	100.0	97.1	98.7	99.1	99.1
NS5	$_{6}$GETLGEKWK$_{14}$	92.0	96.0	98.5	98.8	100.0	100.0	100.0	100.0
	$_{79}$DLGCGRGGWSYY$_{90}$	100.0	100.0	98.5	98.2	100.0	100.0	100.0	100.0
	$_{104}$TKGGPGHEEP$_{113}$	90.8	94.8	98.5	98.8	100.0	100.0	100.0	100.0
	$_{141}$DTLLCDIGESS$_{151}$	100.0	99.4	100.0	99.4	100.0	100.0	100.0	100.0
	$_{209}$PLSRNSTHEMYW$_{220}$	100.0	100.0	100.0	98.8	100.0	100.0	100.0	100.0
	$_{302}$TWAYHGSYE$_{310}$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$	100.0	99.4	98.4	98.7	98.7	99.2	96.3	96.8
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$	100.0	99.4	92.1	93.7	96.1	97.5	100.0	100.0
	$_{468}$AKGSRAIWYMWLGAR$_{482}$	96.6	98.3	100.0	100.0	98.1	98.7	100.0	100.0
	$_{505}$SGVEGEGLH$_{513}$	100.0	100.0	95.3	95.7	98.7	98.8	100.0	100.0
	$_{531}$YADDTAGWDTRIT$_{543}$	97.7	98.9	100.0	100.0	99.4	99.6	100.0	100.0
	$_{568}$IFKLTYQNKVV$_{578}$	100.0	99.4	96.9	95.7	96.9	97.9	100.0	100.0
	$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$	95.4	93.1	83.1	81.0	97.5	98.3	96.3	96.8
	$_{658}$RMAISGDDCVVKP$_{670}$	100.0	100.0	100.0	100.0	95.6	97.1	100.0	100.0
	$_{707}$VPFCSHHFH$_{715}$	97.7	98.9	100.0	100.0	100.0	100.0	96.3	96.8
	$_{765}$LMYFHRRDLRLA$_{776}$	100.0	100.0	98.5	98.8	98.1	98.8	100.0	100.0
	$_{790}$PTSRTTWSIHA$_{800}$	98.9	98.3	98.4	98.8	98.1	98.7	100.0	100.0

[a] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[b] *Rounded to 1 decimal place*

Table 4.3: Distribution and size of the pan-DENV sequences.

DENV protein	Size (aa)	Pan-DENV sequences [a]		
		No.	Size [b]	% of protein [c]
C	113-115	0	0	0
prM	166	0	0	0
E	493-495	2	25	5
NS1	352	7	74	21
NS2a	218	0	0	0
NS2b	130	0	0	0
NS3	618-619	12	122	20
NS4a	150	1	9	6
NS4b	245-249	5	69	28
NS5	900-904	17	215	24
Total	3387-3398	44	514	15

[a] *Sequences of at least nine amino acids that were represented in ≥ 80% of all DENV sequences of each serotype*

[b] *Combined amino acid size of all pan-DENV sequences in the protein*

[c] *Percentage of the combined pan-DENV sequence size over that of the corresponding protein size*

4.3.3 Evolutionary stability of pan-DENV sequences

The evolutionary diversity of each DENV serotype, and the four serotypes combined, was studied by use of Shannon's entropy (Shannon, 1948), modified to examine the variability of nonamer peptide sequences, as described in the Methods (section 4.2.4). The entropy of the proteome of the recorded viruses of each serotype showed numerous long regions of low entropy (≤ 1), reflecting the relatively high degree of intra-serotype sequence conservation, in particular in the NS3, NS4b and NS5 proteins (Figure 4.3 A-D). Of note, however, there were marked differences in the relative degree of entropy of each protein between the four DENV serotypes. For example, NS4b had the least diversity of the proteins of three serotypes, but was

replaced in DENV-2 by NS2b, which was the second most variable in DENV-3. The consequence of the differences in the sequences of each protein between the four serotypes was a marked increase in the peptide entropy across the DENV 1-4 proteomes (Figure 4.3 E), except for 44 sharply defined regions of low nonamer entropy (≤ 0.5) where the sequences were highly conserved in all DENVs (Figure 4.3 E), with no significant difference between the 2005 and 2007 datasets (Table 4.4). The pan-DENV sequences were localized in these 44 regions, with majority of them exhibiting entropy values of ≤ 0.3, corresponding to intra-serotype representation of ≥ 90%. Thus, the congruent consensus- and entropy-based analyses of the DENV nonamer peptides revealed highly conserved and evolutionarily stable pan-DENV sequences distributed in several viral proteins, despite the marked viral diversity defining multiple DENV serotypes, genotypes and variants (Holmes and Burch, 2000). Because of this stability, we predict that the pan-DENV sequences are likely to remain conserved in the future.

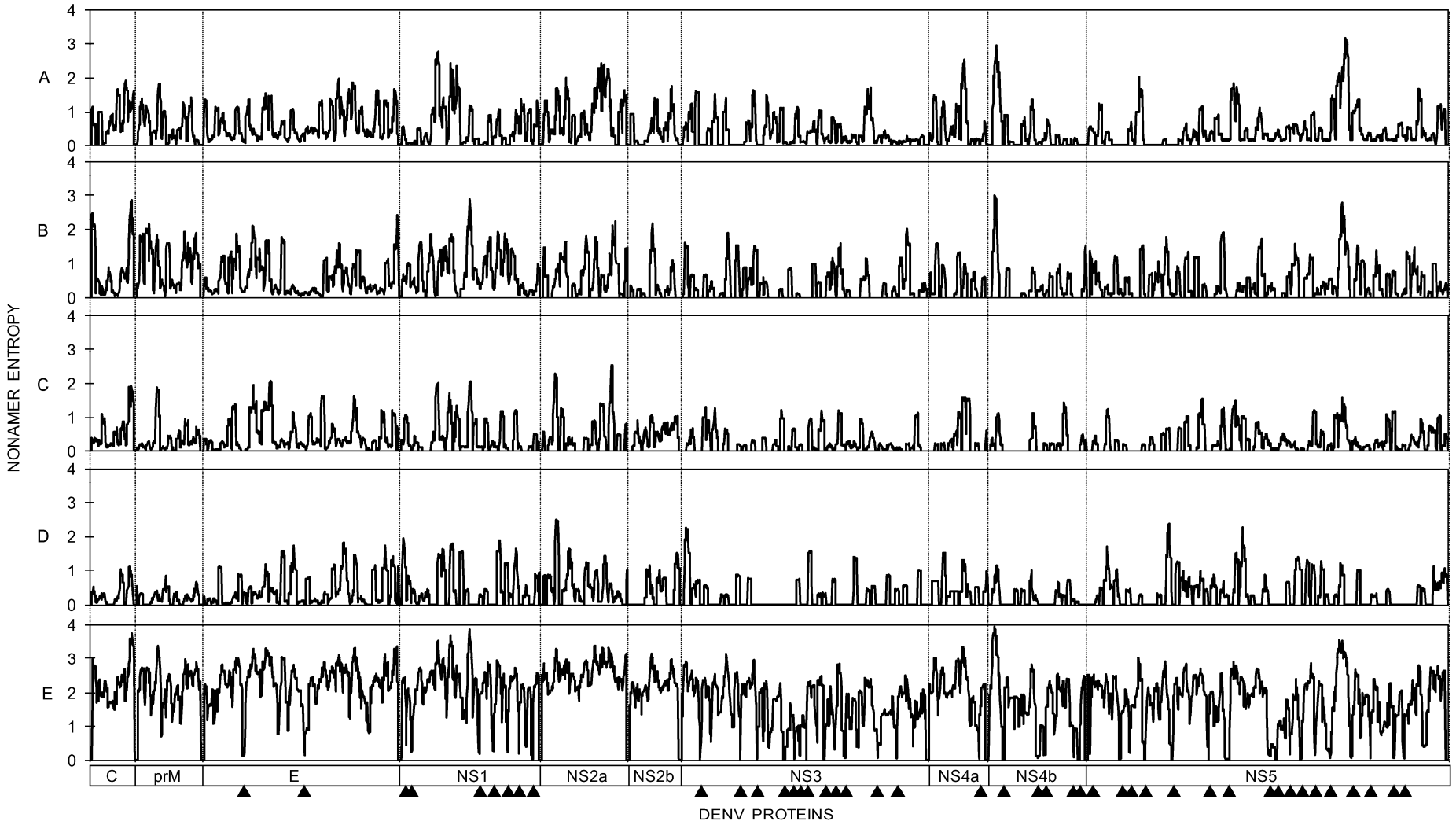

NONAMER ENTROPY
A
B
C
D
E
0
1
2
3
4
C
prM
E
NS1
NS2a
NS2b
NS3
NS4a
NS4b
NS5
DENV PROTEINS

Figure 4.3: Shannon entropy of nonamer peptides within and across DENV serotypes sequences. The entropy values were computed from the alignments of DENV sequences using the Antigenic Variability Analyzer software, as described in Chapter 4.2.4. Values were plotted for DENV-1 (A), DENV-2 (B), DENV-3 (C), DENV-4 (D), and all four DENV serotypes (E) sequences (2005 dataset). Entropy values around protein cleavage sites are non significant, since the corresponding positions cannot be the center of a nonamer (see Chapter 4.2.4). The triangles below indicate the locations of the pan-DENV sequences in the corresponding proteins.

Table 4.4: Pan-DENV sequences, entropy and representation of variants.

DENV protein	Pan-DENV sequence [a]	Peptide entropy [b]		% variant representation [c]	
		2005	2007	2005	2007
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$	0.2	0.2	1-2	1
	$_{252}$VLGSQEGAMH$_{261}$	0.2	0.2	1-2	1-2
NS1	$_{12}$ELKCGSGIF$_{20}$	0.5	0.5	5	5
	$_{25}$VHTWTEQYKFQ$_{35}$	0.3	0.3	3	2-3
	$_{193}$AVHADMGYWIES$_{204}$	0.3	0.3	2-3	2-3
	$_{229}$HTLWSNGVLES$_{239}$	0.3	0.3	3	2-3
	$_{266}$GPWHLGKLE$_{274}$	0.2	0.3	3	3
	$_{294}$RGPSLRTTT$_{302}$	0.2	0.2	3	2
	$_{325}$GEDGCWYGMEIRP$_{337}$	0.2	0.2	< 1-2	< 1-2
NS3	$_{46}$FHTMWHVTRG$_{55}$	0.0	0.0	0	0
	$_{148}$GLYGNGVVT$_{156}$	< 0.1	< 0.1	< 1	< 1
	$_{189}$LTIMDLHPG$_{197}$	0.1	0.1	1	1
	$_{256}$EIVDLMCHATFT$_{267}$	< 0.1	< 0.1	< 1	< 1
	$_{284}$MDEAHFTDP$_{292}$	< 0.1	0.1	< 1	< 1
	$_{296}$AARGYISTRV$_{305}$	0.2	0.2	2	1
	$_{313}$IFMTATPPG$_{321}$	0.0	< 0.1	0	< 1
	$_{357}$GKTVWFVPSIK$_{367}$	0.1	0.2	< 1-1	1-2
	$_{383}$VIQLSRKTFD$_{392}$	0.4	0.3	5	4
	$_{406}$VVTTDISEMGANF$_{418}$	0.2	0.1	1-2	1
	$_{491}$EAKMLLDNI$_{499}$	0.1	0.1	1	1
	$_{537}$LMRRGDLPVWL$_{547}$	0.1	0.2	1	3
NS4a	$_{126}$QRTPQDNQL$_{134}$	0.1	< 0.1	1	< 1
NS4b	$_{35}$PASAWTLYAVATT$_{47}$	0.0	< 0.1	0	< 1
	$_{118}$HYAIIGPGLQAKATREAQKR$_{137}$	0.2	0.2	1-2	1
	$_{139}$AAGIMKNPTVDGI$_{151}$	0.2	0.2	2	1-2
	$_{213}$FWNTTIAVS$_{221}$	0.1	0.1	1	1
	$_{223}$ANIFRGSYLAGAGL$_{236}$	0.1	< 0.1	0-1	0- < 1
NS5	$_{6}$GETLGEKWK$_{14}$	0.2	0.2	2	2
	$_{79}$DLGCGRGGWSYY$_{90}$	0.1	0.1	1	< 1-1
	$_{104}$TKGGPGHEEP$_{113}$	0.2	0.2	3	2
	$_{141}$DTLLCDIGESS$_{151}$	0.0	< 0.1	0	0- < 1
	$_{209}$PLSRNSTHEMYW$_{220}$	0.0	0.1	0	< 1
	$_{302}$TWAYHGSYE$_{310}$	0.0	0.0	0	0
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$	0.1	0.1	0-1	0-1
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$	0.3	0.2	1-3	1-2
	$_{468}$AKGSRAIWYMWLGAR$_{482}$	0.1	0.1	< 1-1	< 1-1
	$_{505}$SGVEGEGLH$_{513}$	0.2	0.2	2	2
	$_{531}$YADDTAGWDTRIT$_{543}$	0.1	0.1	< 1-1	< 1-1
	$_{568}$IFKLTYQNKVV$_{578}$	0.2	0.2	2	2
	$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$	0.4	0.3	1-5	1-5
	$_{658}$RMAISGDDCVVKP$_{670}$	0.2	0.1	1-2	1
	$_{707}$VPFCSHHFH$_{715}$	0.1	0.1	1	1
	$_{765}$LMYFHRRDLRLA$_{776}$	0.1	0.1	1	1
	$_{790}$PTSRTTWSIHA$_{800}$	0.2	0.1	1-2	1

[a] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[b] *Maximum nonamer peptide entropy across all DENV sequences (rounded to 1 decimal place)*

[c] *Minimum and maximum percentage representation of nonamer variants in all DENV sequences (rounded to whole number)*

4.3.4 Representation of nonamer variants in pan-DENV sequences

The combined representation of variant peptides that differed by at least one amino acid from the predominant peptide was analyzed at each nonamer position in the protein alignments. Examples of this analysis for DENV-3 proteins are shown in Table 4.5. Nonamers that lacked entropy (zero entropy) had one sequence in all of the recorded virus isolates, and therefore had no variants. Positions with high entropy can contain many different variant peptides, each at lesser (or equal) frequency than the predominant peptide. The combined representation of variant peptides at each nonamer position across the proteome of each individual DENV serotype was generally low, representing less than 10% of the corresponding sequences, except for some positions where it was more than 50% (Figure 4.4 A-D). Notably, the nonamer position with the highest combined variant representation for each DENV serotype was found in the nonstructural proteins and not the structural ones, with representation values ranging from ~61 to ~78% (DENV-1: NS5, DENV-2: NS5, DENV-3: NS2a, and DENV-4: NS1 and NS3 proteins). When representations of variants across all DENVs were calculated, the majority of all nonamer sites contained variants that together represented ~60-85% of the total DENV sequences at that site (the highest representation of ~85% was in the NS1 protein) (Figure 4.4 E). This was in striking contrast to the 0 to ~5% combined representation of variants at each nonamer position in the pan-DENV sequences, with no significant difference between the 2005 and 2007 datasets (Table 4.4). The majority of all nonamer sites in the pan-DENV sequences lacked variant or contained variants that together represented < 1% of all recorded DENVs. These data further illustrate the extremely high genetic stability of the 44 pan-DENV sequences, among all recorded DENV

sequences and demonstrate that irrespective of the high variability between the sequences of the four DENV serotypes, the representation of variants in the pan-DENV sequences was almost negligible.

Table 4.5: Examples of distribution of variant nonamer peptides in DENV-3.

DENV-3 protein	Nonamer position	No. of sequences	Nonamer peptides [a]	Representation of peptides	Combined % representation of variants [b]	Nonamer entropy [c]
E	14	479	DFVEGLSGA	479 (100%)	0	0
NS2a	176	64	LAGISLLPV	25 (39%)	61	2.4
			LAG**V**SLLPV	11 (17%)		
			LAG**V**SLLP**L**	9 (14%)		
			LA**V**ISLLPV	9 (14%)		
			LAGISLLP**L**	6 (9%)		
			LAGISL**F**PV	2 (3%)		
			LAGISL**M**PV	2 (3%)		
NS4a	86	68	SIGLICVVA	39 (57%)	43	1.5
			SIGLICV**I**A	19 (28%)		
			SIGLICV**IV**	8 (13%)		
			SIGLICV**A**A	2 (3%)		

[a] *The predominant peptide is underlined, the differences are shown in boldface*

[b] *Variants include all the peptides at the position, except the predominant*

[c] *Entropy value of all the peptides at the position (predominant peptide included)*

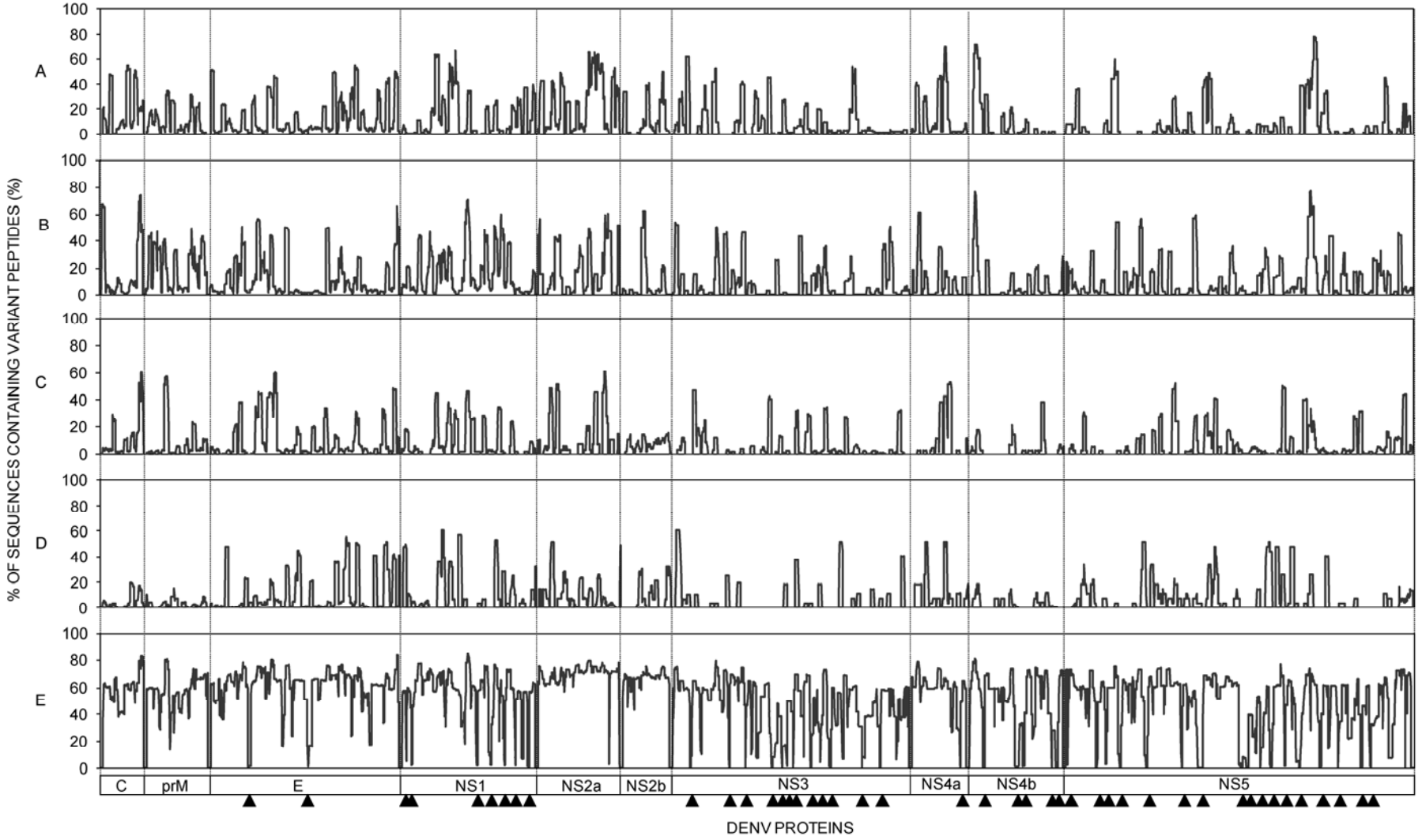

% OF SEQUENCES CONTAINING VARIANT PEPTIDES (%)
A
B
C
D
E
100
80
60
40
20
0
C
prM
E
NS1
NS2a
NS2b
NS3
NS4a
NS4b
NS5
DENV PROTEINS

Figure 4.4: Variant nonamer peptides within and across DENV serotype sequences. The percentage of sequences that contained variant peptides at each nonamer position are shown for DENV-1 (A), DENV-2 (B), DENV-3 (C), DENV-4 (D), and all four DENV serotypes (E) (2005 dataset). Values around protein cleavage sites are non significant (see Figure 4.3). The triangles below indicate the locations of the pan-DENV sequences in the corresponding proteins.

4.3.5 Functional and structural correlates of pan-DENV sequences

Highly conserved protein sequences are likely to represent critical sites and domains (Valdar, 2002). A search of the literature and the Prosite and Pfam databases (Hulo *et al.*, 2006; Bateman *et al.*, 2004) revealed that 27 of the 44 pan-DENV sequences were associated with biological activities (Table 4.6). The two pan-DENV sequences in the E protein correspond to the fusion peptide (positions 98 to 110) and dimerisation domain (Modis *et al.*, 2004; Allison *et al.*, 2001). In NS3, one pan-DENV sequence corresponds to the peptidase family S7 (*Flavivirus* serine protease) domain and comprised the His-51 catalytic residue (Murthy *et al.*, 1999), three sequences correspond to known/putative *Flavivirus* Asp-Glu-Ala-Asp/His (DEAD/H) domain associated with ATP-dependent helicase activity (Xu *et al.*, 2005), and two sequences were predicted to be required for cell attachment and targeting signal for microbodies. In NS5, one pan-DENV sequence corresponds to the conserved methyltransferase (MTase) *S*-adenosyl-L-methionine binding motif I (positons 77-86) involved in viral RNA capping (Egloff *et al.*, 2002), and two sequences correspond to RNA dependent RNA polymerase (RdRp) domain (Yap *et al.*, 2007). Furthermore, six of the 27 pan-DENV sequences were predicted to exhibit post-translational modification(s), including N-glycosylation, protein kinase C and casein kinase II phosphorylation, N-myristoylation and/or amidation (Table 4.6).

It is generally recognized that amino acids buried inside proteins are subject to greater interactions and packing constraints (Haydon and Woolhouse, 1998) than

those exposed on the outer surface. Although none of the DENV protein structures in the PDB (Berman *et al.*, 2000) was full length, 19 of the 44 pan-DENV sequences could be mapped on the available crystallographic models of the E ectodomain (Accession No. 1OAN; 394 out of 493-495 residues), NS3 (1BEF and 2BMF, 181 and 451 out of 618-619 residues, respectively) and NS5 fragments (1R6A, 295 out of 900-904 residues). Eleven of the 19 pan-DENV sequences were buried, two partially exposed and six exposed at the surface of the corresponding structures (Appendix 3). However, these results should be considered preliminary until full length 3-D structures are available.

Table 4.6: Functional and structural properties of pan-DENV sequences.

DENV protein	Pan-DENV sequence [a]	Functional domains and motifs [b]	Putative post-transcriptional modifications [b]
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$	Dimerisation Domain, Fusion Peptide	N-Myristoylation
	$_{252}$VLGSQEGAMH$_{261}$	Dimerisation Domain	-
NS1	$_{12}$ELKCGSGIF$_{20}$	-	N-Myristoylation
	$_{25}$VHTWTEQYKFQ$_{35}$	-	CKII
	$_{294}$RGPSLRTTT$_{302}$	-	PKC
	$_{325}$GEDGCWYGMEIRP$_{337}$	-	N-Myristoylation
NS3	$_{46}$FHTMWHVTRG$_{55}$	Peptidase S7	-
	$_{148}$GLYGNGVVT$_{156}$	-	N-Myristoylation
	$_{189}$LTIMDLHPG$_{197}$	-	CKII
	$_{256}$EIVDLMCHATFT$_{267}$	DEAD/H Domain	-
	$_{284}$MDEAHFTDP$_{292}$	DEAD/H Domain	-
	$_{296}$AARGYISTRV$_{305}$	Microbodies C-Terminal Targeting Signal	PKC
	$_{313}$IFMTATPPG$_{321}$	DEAD/H Domain	-
	$_{357}$GKTVWFVPSIK$_{367}$	-	PKC
	$_{383}$VIQLSRKTFD$_{392}$	-	PKC
	$_{537}$LMRRGDLPVWL$_{547}$	Cell Attachment	-
NS4a	$_{126}$QRTPQDNQL$_{134}$	-	CKII
NS4b	$_{213}$FWNTTIAVS$_{221}$	-	N-Glycosylation
	$_{223}$ANIFRGSYLAGAGL$_{236}$	-	N-Myristoylation
NS5	$_{6}$GETLGEKWK$_{14}$	-	CKII
	$_{79}$DLGCGRGGWSYY$_{90}$	FtsJ-like Methyltransferase Domain	N-Myristoylation
	$_{209}$PLSRNSTHEMYW$_{220}$	-	N-Glycosylation, CKII
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$	-	Amidation
	$_{505}$SGVEGEGLH$_{513}$	-	CKII
	$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$	RdRp Catalytic Domain	N-Myristoylation, CKII
	$_{658}$RMAISGDDCVVKP$_{670}$	RdRp Catalytic Domain	CKII
	$_{790}$PTSRTTWSIHA$_{800}$	-	PKC

[a] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[b] *Described in the literature and/or identified using the Prosite (Hulo et al., 2006) and Pfam (Bateman et al., 2004) databases. Prosite (PS) and Pfam (PF) accession numbers: PS00001, N-Glycosylation; PS00005, Protein Kinase II Phosphorylation (PKC); PS00006, Casein Kinase II Phosphorylation (CKII); PS00008, N-Myristoylation; PS00009, Amidation; PS00016, Cell Attachment; PS00342, Microbodies C-terminal Targeting Signal; PS50507, RNA-dependent RNA polymerase (RdRp) Catalytic Domain; PF00869, Dimerisation Domain; PF00949, Peptidase S7; PF01728, FtsJ-like Methyltransferase; PF07652, Flavivirus DEAD/H Domain.*

4.3.6 Distribution of pan-DENV sequences in nature

Twenty-seven (27) of the 44 pan-DENV sequences overlapped at least nine amino acid sequences of as many as 64 other viruses of the family *Flaviviridae,* genus *Flavivirus* (Figure 4.5). Zika virus shared 22 of the 27 sequences; Ilheus and Kedougou viruses, 18; and representatives of some of the significant human pathogens, West Nile, St. Louis encephalitis, Japanese encephalitis, Yellow fever and Tick-borne encephalitis viruses, shared from 16 to 9 pan-DENV sequences. Thirteen (13) of the 27 sequences represented NS5, of which nine were present in at least 27 *Flavivirus* species; nine represented NS3, of which two were found in 35 and 23 species; one E sequence was found in 19 species; and the remaining were in NS1 and NS4b (Figure 4.6; Table 4.7). Five (5) of the 27 were associated with known biological activities ($NS5_{79\text{-}90}$ MTase, $NS5_{658\text{-}670}$ RdRp, $NS3_{46\text{-}55}$ peptidase S7, $NS3_{284\text{-}292}$ DEAD/H and $E_{97\text{-}111}$ dimerisation/fusion domains). Interestingly, two sequences, $NS3_{406\text{-}418}$ and $NS5_{597\text{-}616}$, overlapped nine amino acid sequences of the cell fusing agent virus polyprotein-like protein from the mosquito *Aedes albopictus* (Crochu *et al.*, 2004), and the phage-related tail fibre protein-like protein from the bacterium *Chromohalobacter salexigens DSM 3043*, respectively.

The representation of the pan-DENV sequences ranged from high to low across reported sequences of several of the highly studied flaviviruses (Table 4.7): St.

Louis encephalitis, West Nile, Japanese encephalitis, Murray Valley encephalitis, Usutu, Kokobera, Ilheus, Tick-borne encephalitis, Langat, Omsk hemorrhagic fever, Louping ill, Powassan, Kyasanur forest disease and Yellow fever viruses. Protein sequence data for the rest of the flaviviruses that shared pan-DENV sequences was limited (< 10 sequences) in the public database.

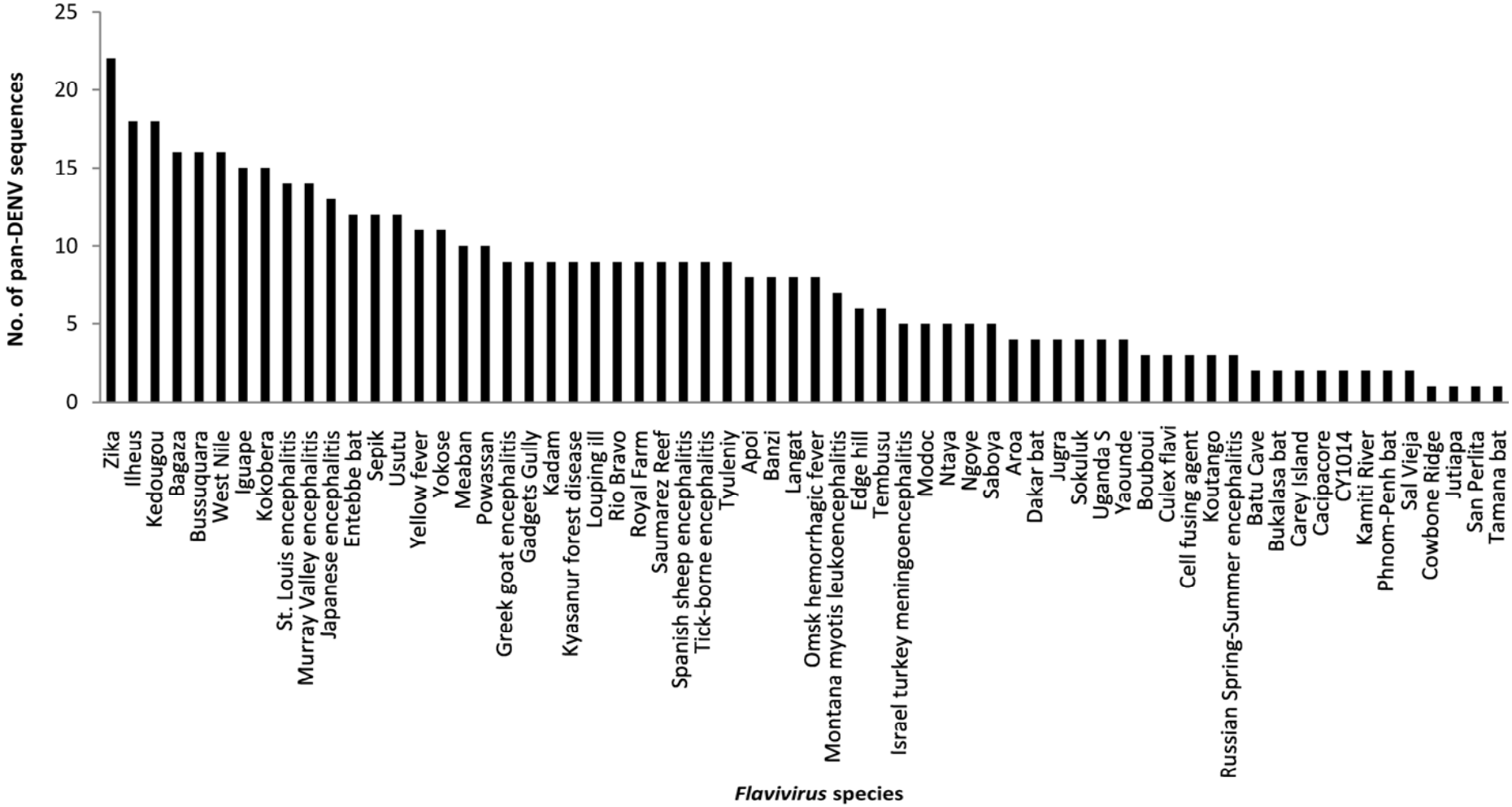


Figure 4.5: Number of pan-DENV sequences conserved in other flaviviruses.

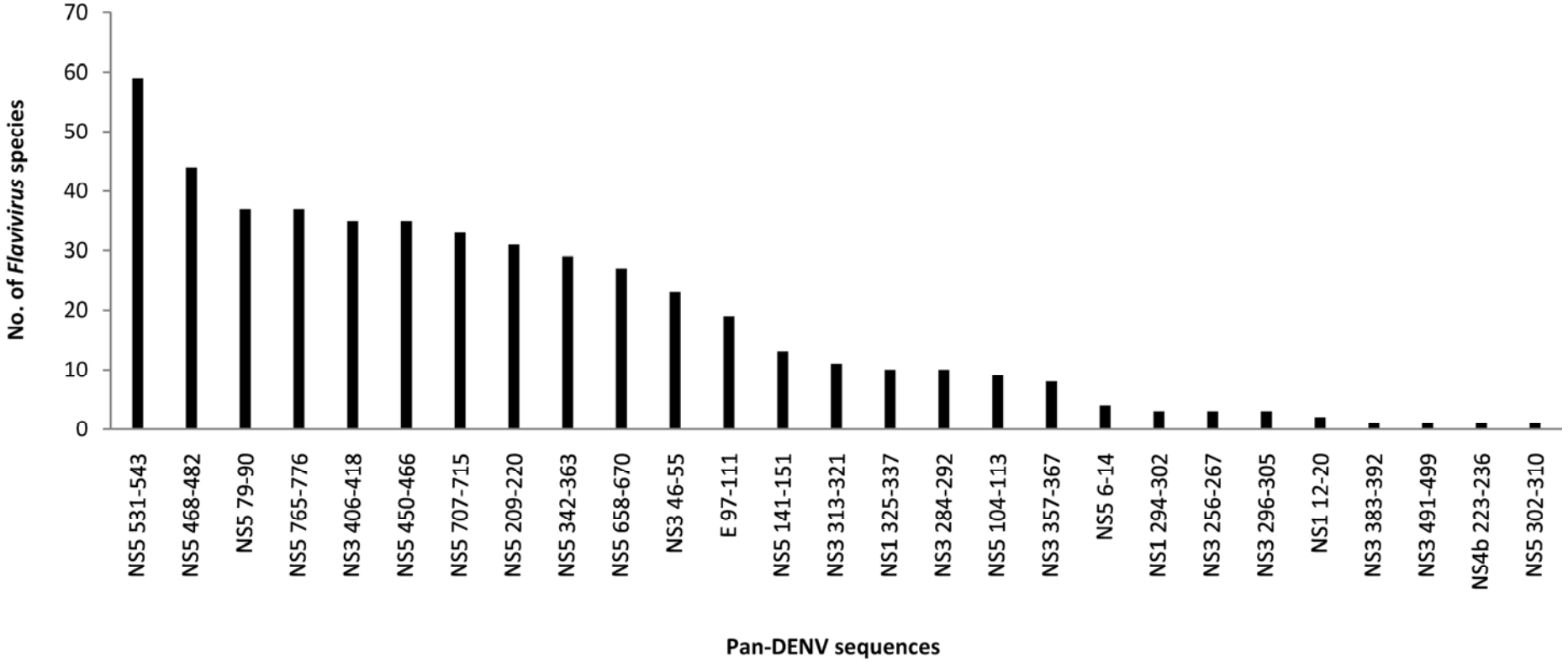


Figure 4.6: Number of other flaviviruses sharing the Pan-DENV sequences.

Table 4.7: Distribution of pan-DENV sequences in other flaviviruses.

DENV protein	Pan-DENV sequence [a]	Species (#) [b]	Percentage representation (%) | number of sequence analyzed [c]													
			LEV	WNV	JEV	MVE	UV	KBV	IH	TBEV	LV	OMSK	LIV	PV	KFDV	YFV
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$	19	97|77	97|276	97|250											82|179
NS1	$_{12}$ELKCGSGIF$_{20}$	2														
	$_{294}$RGPSLRTTT$_{302}$	3			3|58											
	$_{325}$GEDGCWYGMEIRP$_{337}$	10	90|30	96|138	31|186											
NS3	$_{46}$FHTMWHVTRG$_{55}$	23								89|19						87|23
	$_{256}$EIVDLMCHATFT$_{267}$	3														
	$_{284}$MDEAHFTDP$_{292}$	10		99|134	98|54											
	$_{296}$AARGYISTRV$_{305}$	3	9|98	4|137												
	$_{313}$IFMTATPPG$_{321}$	11	96|26	100|134	98|53											
	$_{357}$GKTVWFVPSIK$_{367}$	8		94|141					17|12							
	$_{383}$VIQLSRKTFD$_{392}$	1														
	$_{406}$VVTTDISEMGANF$_{418}$	35	34|77	91|146	21|248					68|25						
	$_{491}$EAKMLLDNI$_{499}$	1														
NS4b	$_{223}$ANIFRGSYLAGAGL$_{236}$	1														
NS5	$_{6}$GETLGEKWK$_{14}$	4														
	$_{79}$DLGCGRGGWSYY$_{90}$	37	77|35	96|140	20|244					82|38				83|18	40|10	78|27
	$_{104}$TKGGPGHEEP$_{113}$	9	93|28	89|148												
	$_{141}$DTLLCDIGESS$_{151}$	13		100|134	< 1|268											84|25
	$_{209}$PLSRNSTHEMYW$_{220}$	31	90|29	100|134						61|28				100|17		95|22
	$_{302}$TWAYHGSYE$_{310}$	1														
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$	29	33|79	50|272	18|289	18|17	23|13		25|12	32|34				60|45		51|41
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$	35	24|103	33|344	69|74	20|15		7|14	14|14	63|27			11|19	41|41		23|90
	$_{468}$AKGSRAIWYMWLGAR$_{482}$	44	32|19						62|13	82|22	42|12			41|41		11|208
	$_{531}$YADDTAGWDTRIT$_{543}$	59	90|30	73|193	88|58				80|10	69|26		93|14		84|19		85|26
	$_{658}$RMAISGDDCVVKP$_{670}$	27	73|37	50|272	95|55	50|10								47|34		3|37
	$_{707}$VPFCSHHFH$_{715}$	33								100|19		100|13		76|17		95|22
	$_{765}$LMYFHRRDLRLA$_{776}$	37	64|42	91|151	83|59				44|18							

[a] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[b] *Species (#) column indicates the number of viral species that shared at least nine consecutive amino acids of the pan-DENV sequence*

[c] *Percentage representation (rounded to whole number) of the pan-DENV sequences is only shown for viral species with ≥ 10 total sequences reported at the NCBI Entrez Protein database. These viral species included: LEV, St. Louis encephalitis virus; WNV, West Nile virus; JEV, Japanese encephalitis virus; MVE, Murray Valley encephalitis virus; UV, Usutu virus; KBV, Kokobera virus; IH, Ilheus virus; TBEV, Tick-born encephalitis virus; LV, Langat virus; OMSK, Omsk hemorrhagic fever virus; LIV, Louping ill virus; PV, Powassan virus; KDFV, Kyasanur forest disease virus; and YFV, Yellow fever virus. However, despite having a total of ≥ 10 sequences reported, some of these viruses had less than 10 of the relevant conserved sequence (indicated by cells shaded in grey). Empty cells indicate no match between the pan-DENV sequences and the Flavivirus.*

4.3.7 Known and predicted HLA supertype-restricted, pan-DENV T-cell epitopes

Literature survey and database search revealed that 10 of the pan-DENV sequences (9 in NS3, one in E) overlapped by nine or more amino acids 15 previously reported DENV T-cell epitopes immunogenic in human. Their HLA restriction, when known, showed both class II (DR*15, DPw2) and class I (A*11) specificities (Table 4.8). Further evaluation of the immune-relevance of the pan-DENV sequences included a search for candidate putative promiscuous HLA supertype-restricted T-cell epitopes within these regions by use of several computational algorithms: NetCTL (Larsen *et al.*, 2005), Multipred (Zhang *et al.*, 2005b), ARB (Bui *et al.*, 2005) and TEPITOPE (Bian and Hammer, 2004). Overall, 34 of 44 (~77%) pan-DENV sequences (Figure 4.7), identified in the NS5, NS3, NS1, E and NS4a proteins were predicted to contain 100 supertype-restricted binding nonamers (Appendix 4). The majority (88/100) of the predicted promiscuous HLA-binding nonamers were present in ≥ 95% of the sequences of each DENV serotype (Appendix 5). Thirty-one (~91%) of the 34 putative supertype pan-DENV sequences contain HLA-binding nonamers for multiple HLA supertypes. Clusters (hotspots) of two or more overlapping HLA-binder

nonamer core peptides were present in 27 (~79%) of the 34 putative supertype pan-DENV sequences. About half (14/27) of these clusters contain three or more nonamer binders overlapping by eight amino acids, covering most or the entire corresponding conserved region.

Table 4.8: Human T-cell epitopes within the pan-DENV sequences.

DENV protein	Pan-DENV sequence [a]	Immunogenic T-cell epitopes [b]			
		Sequence [c]	T subset	HLA Ag	Reference(s) [d]
E	$_{252}$VLGSQEGAMH$_{261}$	KKQDVVVLGSQEGAM	-	-	1
NS3	$_{46}$FHTMWHVTRG$_{55}$	TFHTMWHVTRGAVLM	CD4	-	1
	$_{148}$GLYGNGVVT$_{156}$	KVVGLYGNGVVTRSG	CD4	DR*15	1
	$_{189}$LTIMDLHPG$_{197}$	KRLTIMDLHPGAGKT	CD4	-	2
		RKLTIMDLHPGSGKT	CD4	-	2
		RKLTIMDLHPGAGKT	CD4	-	2
		RNLTIMDLHPGSGKT	CD4	-	2
	$_{256}$EIVDLMCHATFT$_{267}$	EHTGREIVDLMCHAT	CD4	-	1
		EIVDLMCHATFTMRL	CD4	-	1
		EIVDLMCHAT	CD4	DPw2	3,4
	$_{284}$MDEAHFTDP$_{292}$	LIIMDEAHFTDPASI	-	-	1
	$_{313}$IFMTATPPG$_{321}$	AGIFMTATPPGSRDP	-	-	1
	$_{357}$GKTVWFVPSIK$_{367}$	TVWFVPSIK	CD8	A*11	5
	$_{383}$VIQLSRKTFD$_{392}$	KKVIQLSRKTFDSEY	-	-	1
	$_{406}$VVTTDISEMGANF$_{418}$	NDWDFVVTTDISEMG	-	-	1

[a] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[b] *Dashes, not determined*

[c] *Sequences present in the pan-DENV sequences are underlined*

[d] *1 - (Simmons et al., 2005); 2 - (Mangada and Rothman, 2005); 3- (Kurane et al., 1993); 4 - (Okamoto et al., 1998); 5 - (Loke et al., 2001)*

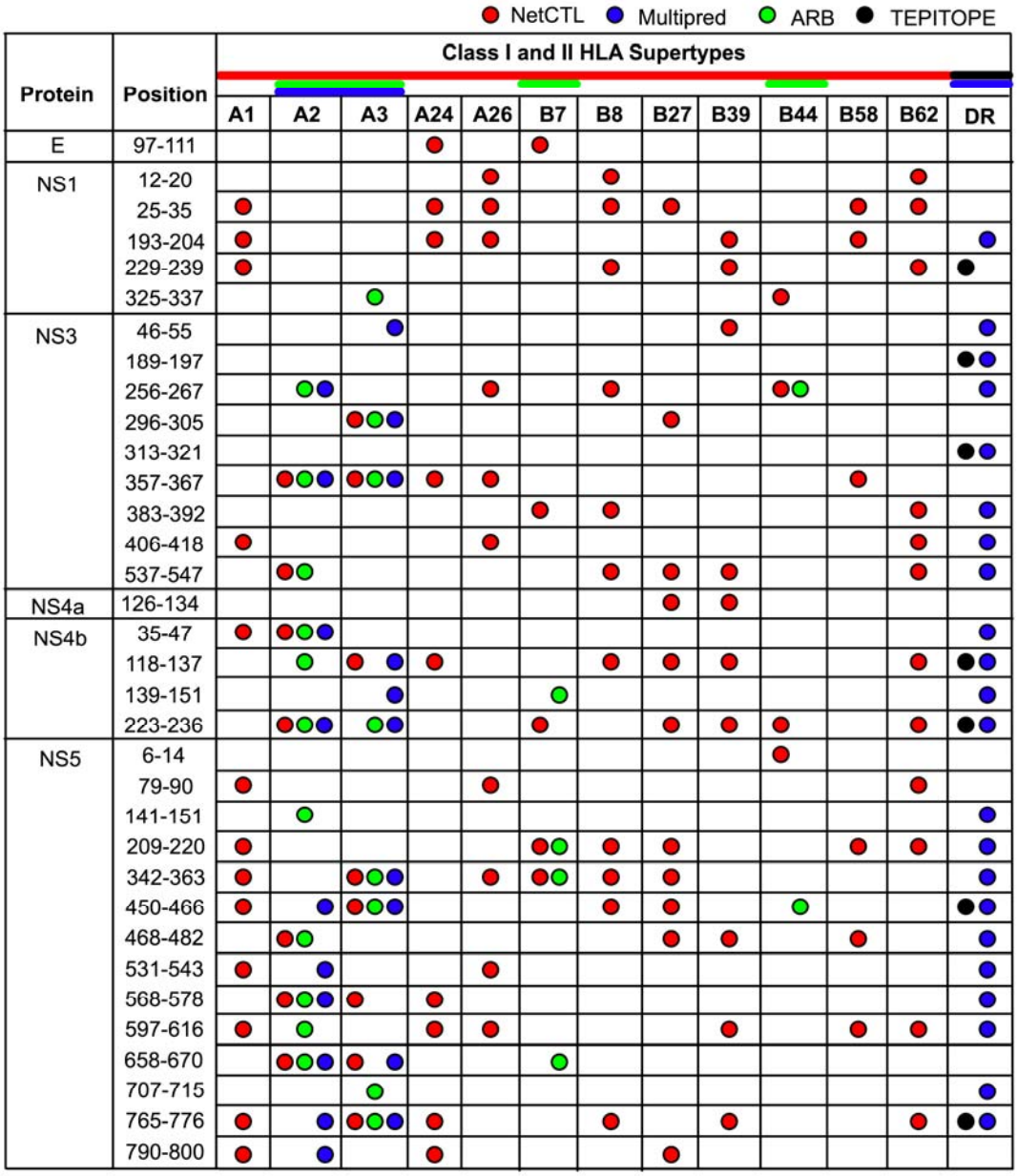


Figure 4.7: Putative HLA supertype-restricted, pan-DENV T-cell epitopes pre-screened by computational algorithms. Amino acid positions of the pan-DENV sequences are numbered according to the sequence alignments of the four DENV serotypes; the corresponding DENV proteins are indicated on the left. Predicted HLA-restricted T-cell epitopes were identified using NetCTL, Multipred, ARB and TEPITOPE algorithms (see Chapter 4.2.8).

4.3.8 Immunogenicity of HLA-DR-restricted pan-DENV sequences in HLA transgenic mice

The immunogenicity of the pan-DENV sequences was also analyzed by assays of peptide-specific HLA-restricted T-cell responses in murine H-2 class II-deficient, HLA-DR transgenic mice expressing three prototypic HLA-DR alleles, corresponding to the divergent subgroups HLA-DR2 (DRB1*1501), HLA-DR3 (DRB1*0301), and HLA-DR4 (DRB1*0401). Mice were immunized with pools of overlapping peptides covering the sequences of the E, NS1, NS3, and NS5 proteins of DENV-3, and HLA-DR-restricted CD4 T-cell responses were assessed by IFN-γ ELISpot assays using CD8-depleted splenocytes. Thirty peptides eliciting positive T-cell responses in the HLA transgenic mice contain nine or more consecutive amino acids of 22 pan-DENV sequences, that were localized in the NS5 (11), NS3 (6), NS1 (4), and E proteins (one) (Table 4.9). Overall, 9, 10 and 18 peptides elicited positive responses in HLA-DR2, -DR3, and/or -DR4 transgenic mice, respectively; 20 correspond to sequences of NS5, 10 of NS3, six of NS1, and one of E. Furthermore, at least seven of the pan-DENV sequences, all localized in the NS5 and NS1 proteins, contain promiscuous T-cell epitopes for multiple HLA-DR alleles (Table 4.9). These data, together with those previously reported (Table 4.8)**,** showed that 26 of the 44 pan-DENV sequences, distributed predominantly in the NS5 and NS3 proteins, and to a lesser extent in NS1 and E, contain numerous HLA-restricted class II and/or class I epitopes demonstrated by assays of T-cell responses *in vivo*.

Table 4.9: Immunogenicity of the pan-DENV sequences in HLA-DR transgenic mice.

DENV protein	Pan-DENV sequence [b]	Ag-specific CD4 T-cell responses [a]			
		Peptide sequences (DENV-3) [c]	IFN-γ-SFC/10^6splenocytes ± SD [d]		
			DR2	DR3	DR4
E	$_{252}$VLGSQEGAMH$_{261}$	PEVVVLGSQEGAMHT	-	-	88 ± 34
NS1	$_{193}$AVHADMGYWIES$_{204}$	AVHADMGYWIESQKN	-	17 ± 1	-
	$_{229}$HTLWSNGVLES$_{239}$	WPKSHTLWSNGVLES	-	129 ± 3*	-
		HTLWSNGVLESDMII	-	131 ± 103	37 ± 3
	$_{266}$GPWHLGKLE$_{274}$	HTQTAGPWHLGKLE	-	333 ± 6	-
	$_{294}$RGPSLRTTT$_{302}$	TRGPSLRTTTVSGKL	-	-	11 ± 4
NS3	$_{189}$LTIMDLHPG$_{197}$	KKRNLTIMDLHPGSG	-	-	50 ± 16
	$_{296}$AARGYISTRV$_{305}$	ASIAARGYISTRVGM	40 ± 14	-	-
		ARGYISTRVGMGEAA	9 ± 4	-	-
	$_{313}$IFMTATPPG$_{321}$	EAAAIFMTATPPGTA	-	-	474 ± 116
		IFMTATPPGTADAFP	-	-	323 ± 287
	$_{357}$GKTVWFVPSIK$_{367}$	TDFAGKTVWFVPSIK	48 ± 15	-	-
		GKTVWFVPSIKAGND	396 ± 14	-	-
	$_{383}$VIQLSRKTFD$_{392}$	KKVIQLSRKTFDTEY	-	21 ± 3	-
	$_{406}$VVTTDISEMGANF$_{418}$	FVVTTDISEMGANFK	-	-	408 ± 104
		TDISEMGANFKADRV	-	152 ± 33	-
NS5	$_{302}$TWAYHGSYE$_{310}$	DENPYKTWAYHGSYEVK	126 ± 10*	-	14 ± 5
		TWAYHGSYEVKATGSA	161 ± 20*	-	63 ± 17
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$	MVTQMAMTDTTPFGQQR	-	-	28 ± 0*

$_{450}$CVYNMMGKREKKLGEFG$_{466}$	GSCVYNMMGKREKKLGE	-	-	13 ± 2
$_{505}$SGVEGEGLH$_{513}$	NSYSGVEGEGLHKLGYI	-	-	184 ± 15
$_{531}$YADDTAGWDTRIT$_{543}$	KIPGGAMYADDTAGWDT	-	-	46 ± 3
$_{568}$IFKLTYQNKVV$_{578}$	ANAIFKLTYQNKVVKVQ	577 ± 384	-	24 ± 9*
$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$	VMDIISRKDQRGSGQVG	-	88 ± 1	-
$_{658}$RMAISGDDCVVKP$_{670}$	VERLKRMAISGDDCVVK	-	159 ± 24	16 ± 6
	MAISGDDCVVKPIDDRF	-	249 ± 39	-
$_{707}$VPFCSHHFH$_{715}$	DWQQVPFCSHHFHELIM	32 ± 8*	34 ± 11	-
$_{765}$LMYFHRRDLRLA$_{776}$	MYFHRRDLRLASNAI	75 ± 16*	-	33 ± 9
$_{790}$PTSRTTWSIHA$_{800}$	VHWVPTSRTTWSIHAHH	-	-	83 ± 1
	SRTTWSIHAHHQWMTTE	-	-	122 ± 46

[a] *Assessed by IFN-γ ELISpot assay in HLA-DR2 (DRB1*1501), HLA-DR3 (DRB1*0301) and HLA-DR4 (DRB1*0401) transgenic mice immunized with DENV-3 peptides (see Chapter 4.2.8)*

[b] *Amino acid positions numbered according to the sequence alignments of the four DENV serotypes*

[c] *Sequences present in the pan-DENV sequences are underlined*

[d] *SFC, spot-forming cells; SD, standard deviation. Representative results from at least 2 immunized transgenic mice are shown, except when indicated by an asterisk (*)*

4.4 Discussion

In this study, the author identified and characterized pan-DENV sequences that were highly conserved in the majority of the reported sequences of each serotype. The author regards all the 44 pan-DENV sequences as potential candidate DENV *PE*s even though only 34 were predicted to contain promiscuous T-cell epitopes, as there is a possibility of the remaining 10 to also be immunologically relevant in the context of the HLA supertypes predicted or others when experimentally validated. The characterization of the 44 elucidated pan-DENV sequences shed several insights into the nature of these sequences and their multiple potential applications.

The large number of sequences analyzed (12,404 as of December 2007), and their wide spatial and temporal (1945-2007; based on all DENV records with available annotation - data not shown) distribution, offered information for a broad survey of DENV diversity in nature. The significant sequence variations between the proteins of the four DENV serotypes represent a cardinal issue for the development of a tetravalent DENV vaccine that provides robust protection against each DENV serotype. Subtle amino acid substitutions within T-cell epitopes restricted by a given HLA allomorph, such as in the event of sequential heterologous infections, or between a vaccine formulation and a subsequent natural infection (Rothman, 2004), can dramatically alter the phenotype of the specific T cells, resulting in a wide range of effects from agonism to antagonism (Nishimura *et al.*, 2004; Kalergis and Nathenson, 2000; Madrenas and Germain, 1996; Sloan-Lancaster and Allen, 1996; Evavold *et al.*, 1993). Because of the extent of intra-serotype (1 to 21%) and inter-serotype (14 to 67%) amino acid variability among DENV isolates (Chapter 3; (Khan *et al.*, 2006a)), many nonamer T-cell epitopes contain single or multiple amino acid difference(s). When the four DENV serotypes were analyzed together, a majority of

the nonamer positions across the proteome exhibited variants that together were present in ~60 to ~85% of all sequences. The frequencies of variant peptides across the four DENV serotypes suggest that vaccine strategies incorporating whole DENV immunogens, such as inactivated and recombinant subunit vaccines, live attenuated viruses, or chimeric viruses expressing structural DENV genes, have potential to elicit T-cell responses to altered peptide ligands. This phenomenon is also likely to occur in individuals exposed to several flaviviruses, such as DENV, JEV and YFV that are co-circulating in regions of Asia, India or South America, or following vaccination (Moran *et al.*, 2008).

While the immune correlates of DENV protection remain poorly documented, there is evidence that both neutralizing antibody and specific T-cell mediated responses are required (Whitehead *et al.*, 2007; Rothman, 2004). The incorporation of defined HLA-restricted T-cell epitopes within DENV vaccine candidates might improve vaccine efficiency by increasing T-cell help to sustain a robust, long-lived immunity, and possibly through direct cytostatic and cytotoxic effects on infected cells. For tetravalent formulations, it may be relevant to focus primarily on sequences that are conserved in all four DENV serotypes and to avoid the regions of T-cell immunity that are highly variable, unless they are strictly serotype-specific and intra-serotype conserved (Mangada and Rothman, 2005; Mongkolsapaya *et al.*, 2003). An additional criterion for the selection of T-cell targets is the need for epitopes with broad HLA representation, as it has been emphasized in the recognition of HLA supertypes (Sette *et al.*, 2001; Sette and Sidney, 1999). The 44 pan-DENV sequences of at least 9 aa, covering 514 aa or about 15% of the complete DENV polyprotein of ~3390 aa, and conserved in at least 80% of all recorded DENV sequences (34 of the 44 (~77%) were conserved in ≥ 95% of DENV sequences) are attractive candidates

for this purpose as they satisfy these selection criteria: i) highly conserved across all four serotypes, ii) low variant representation iii) broad HLA coverage.

Further, identified conserved sequences have shown remarkable stability over the entire history of DENV sequences deposited in the NCBI Entrez Protein database, as illustrated by their low peptide entropy values. In addition, 27 of the pan-DENV sequences were conserved in 64 other flaviviruses, as further evidence of prolonged evolutionary stability within this genus, as previously discussed by others (Billoir *et al.*, 2000; Kuno *et al.*, 1998; Henchal and Putnak, 1990). Interestingly, two of the pan-DENV sequences are also present in the proteomes of non-viruses, such as *Aedes albopictus* mosquito and the bacterium *Chromohalobacter salexigens*, possibly because of genetic recombination between phyla (Crochu *et al.*, 2004). It is likely that the pan-DENV sequences have been under selection pressure to fulfill critical biological and/or structural properties, some of which have been identified for the E (fusion peptide, dimerization domain), NS3 (peptidase S7, DEAD/H domains) and NS5 proteins (MTPase, RdRp domains) (Yap *et al.*, 2007; Xu *et al.*, 2005; Modis *et al.*, 2004; Egloff *et al.*, 2002; Allison *et al.*, 2001; Murthy *et al.*, 1999). Hence, these conserved sequences are unlikely to significantly diverge in newly emerging DENV isolates in the future, and thus further support their utility as targets for the development of specific anti-viral compounds and vaccine candidates.

Although the conservation of the pan-DENV sequences to other flaviviruses suggests that they are likely to be conserved in the future due to their structural/functional importance, it also posits that potential variants can originate from these flaviviruses, following co-infection or vaccination and secondary infection. The analysis of distribution of *PE*s of a virus of interest (target pathogen) in other viruses and organisms, therefore, provides a platform to assess the potential risk

of altered peptide ligands to *PE*s of the target pathogen, resulting from sequence conservation to other viruses. *PE*s of a target pathogen also found in other viruses, but with a low representation, have a high potential for altered peptide ligand effect due to the large number of potential variants that can be contributed by these other viruses. Further, it should be noted that variants can also originate from viruses that are not from the same genus as the target pathogen. T-cells in nature have been observed to cross-react with different viral species (heterologous immune responses) (Welsh and Fujinami, 2007). For example, T-cells specific to distinct Influenza A virus antigens cross-reacted with Hepatitis C virus (HCV) (Wedemeyer *et al.*, 2001), HIV (Acierno *et al.*, 2003) and Epstein-Barr virus (EBV) (Clute *et al.*, 2005) antigens. Thus, while the consequences of such extensive possible cross-reactive immunity are hypothetical, the author proposes, for vaccine formulation, that it is prudent to select *PE*s that are specific to the pathogen and, thus, representative of a minimal number of variant sequences across other viral species. By selecting sequences with minimized variant representation for vaccine formulation, the likelihood that a vaccinated individual, in their lifetime, will be exposed to variant determinant sequences, either from the target pathogen or other viruses, is greatly reduced. Rational selection of such sequences is critical because factors, such as globalization, increase in travel and increase in life expectancy, are continuously increasing the risk of infection by multiple viruses.

A number of the pan-DENV sequences were predicted to be promiscuous to multiple HLA supertypes, in addition to multiple alleles of a given HLA supertype. Such a degree of promiscuity has previously been observed by others for DENV (Gagnon *et al.*, 1996) and HIV peptides (Wilson *et al.*, 2003), among others. The existence of conserved T-cell epitopes specific for multiple HLA supertypes further supports their potential as vaccine targets, since they would provide broader

population coverage (Wilson *et al.*, 2003). Many of the predicted HLA binding nonamers were localized in clusters, as has also observed in HLA transgenic mice immunized with WNV proteins and DNA encoding the SARS coronavirus N protein (Gupta *et al.*, 2006), and has been reported in studies of HIV serotype 1 proteins (Brown *et al.*, 2003; Surman *et al.*, 2001; Shankar *et al.*, 1996; Berzofsky *et al.*, 1991), the outer membrane protein of *Chlamydia trachomatis* (Kim and DeMars, 2001), and other antigens (Gupta *et al.*, 2006).

The global approach described herein provides a framework and methodology for large-scale and systematic analysis of *PE*s of other pathogens, in particular for rapidly evolving viruses such as Influenza A virus (Heiny *et al.*, 2007) and HIV (Wilson *et al.*, 2003). These studies will offer insights into their diversity and evolutionary history, together with providing critical data for rational vaccine development, structure-based design of candidate inhibitory compounds, and improvement of the current diagnostic methods (Leyssen *et al.*, 2000).

4.5 Chapter summary

***Background*:** Short, conserved viral sequence fragments that contain promiscuous T-cell epitopes are attractive candidates to cover antigenic diversity and for vaccine formulation. Therefore, the author analysed all available DENV sequence data in public databases to identify and characterize peptides that cover antigenic diversity (*PE*s) of the virus: sequence regions conserved across sequences of the four DENV serotypes (pan-DENV sequences) and are immunologically relevant in the context of HLA supertypes.

***Results*:** A large-scale identification and analysis of evolutionarily highly conserved amino acid sequences of the entire DENV proteome, with a focus on

sequences of nine amino acids or more, and thus immune-relevant as potential T-cell epitopes was undertaken. DENV protein sequence data were collected from the NCBI Entrez Protein database in 2005 (9,512 sequences) and again in 2007 (12,404 sequences). Forty-four (44) sequences (pan-DENV sequences), mainly those of nonstructural proteins and representing ~15% of the DENV polyprotein length, were identical in 80% or more of all recorded DENV sequences. Of these 44 sequences, thirty-four (~77%) were present in ≥ 95% of sequences of each DENV serotype, and 27 (~61%) were conserved in other flaviviruses. The frequencies of variants of the pan-DENV sequences were low (0 to ~5%), as compared to variant frequencies of ~60 to ~85% in the non pan-DENV sequence regions. The majority of the conserved sequences were shown to be immunologically relevant: 34 contain numerous predicted HLA supertype-restricted peptide sequences, and 26 contain T-cell epitopes identified by studies with HLA-transgenic mice and/or reported to be immunogenic in humans.

Conclusions: The author identified and characterized 44 pan-DENV sequences as potential DENV candidate *PE*s. The conservation of these sequences through the entire recorded DENV genetic history supports their possible value for diagnosis, prophylactic and/or therapeutic applications. The combination of bioinformatics and experimental approaches applied herein provides a framework for large-scale and systematic analysis of *PE*s of other pathogens, in particular, for rapidly mutating viruses, such as Influenza A virus and HIV.

Chapter 5 A Systematic Bioinformatics Pipeline for Rational Selection of Vaccine Candidates Targeting Antigenic Diversity

5.1 Introduction

New developments in bioinformatics and other computational methodologies, combined with the broad versatility in the design and synthesis of genetic (DNA) vaccines, underlay new strategies for the novel design of antigen-specific, peptide-based vaccines against the many pathogens, such as HIV and influenza, that have been observed to be resistant to conventional vaccine therapy (Sette and Fikes, 2003; Sette *et al.*, 2001). Early clinical trials of peptide-based vaccines for HIV, malaria and tuberculosis have produced promising results (Robinson and Amara, 2005; Wilson *et al.*, 2003), supporting the protective and therapeutic uses of these vaccines. T-cell epitopes, important for cytolytic and regulatory responses to pathogens (Pulendran and Ahmed, 2006; Zinkernagel and Hengartner, 2004; Esser *et al.*, 2003), are necessary elements of these vaccines. The rational selection of protein antigen sequences that function as T-cell epitopes in vaccine formulations is therefore crucial for successful application of this vaccination strategy (Brusic and August, 2004; Sette *et al.*, 2001).

This selection of pathogen antigen sequences to be included in peptide-based vaccines must address several determinative issues. The goal is to identify relevant T-cell epitopes, both HLA class I and II that are both effective and sufficient in vaccine protection against pathogen challenge. A major question is the degree of protection that can be achieved without the concomitant administration of neutralizing antibody epitopes. Vaccines must also protect a broad spectrum of human population against as wide a variety of pathogenic strains as possible. It is therefore important to choose epitopes that cover antigenic diversity: the diversity of both the pathogen and the host; short sequence fragments of the antigen that are conserved in all or majority of the pathogen isolates and variants and contain high concentration of promiscuous T-cell

epitopes that bind to several alleles of HLA supertypes for maximal population coverage (Sette and Sidney, 1999) satisfy these requirements. The bioinformatics-based approaches serve as a means to enhance the optimal selection of potential targets of immune response followed by experimental validation, typically by testing these antigen sequences in immunological assays. In this chapter, the author describes a combined bioinformatics and molecular strategy for vaccine development focusing on covering antigenic diversity (identification and characterization of *PE*s).

5.2 Framework for rational selection of peptide-based vaccine targets that cover antigenic diversity

5.2.1 Data collection and preparation

Predictions about future mutations are derived from past the evolutionary history. It is therefore important to collect sequences that are as representative as possible of the genetic variants of the pathogen, over extended periods of time and broad geographical ranges. Ideally, all available protein sequences pertaining to the pathogen should be collected from major public databases, such as the NCBI Entrez Protein database (www.ncbi.nlm.nih.gov/entrez). Since public databases often contain errors and discrepancies, a data cleaning process is needed to correct such anomalies (Khan *et al.*, 2006a; Khan *et al.*, 2006b; Srinivasan *et al.*, 2002). For example, annotation errors and discrepancies in 17 DV records were identified and corrected prior to analysis (Appendix 2). While several methods are available, the author found our in-house ABK structural rule-based approach (Miotto *et al.*, 2005) well suited to this type of task, allowing fully annotated sets of over 40,000 influenza protein sequences to be cleaned and independently verified in two weeks. The cleaned dataset is then grouped according to established classification, such as based on

genotype/serotype/clade, and the resulting grouped datasets are then further sub-grouped according to the proteins coded by the pathogen to facilitate systematic analysis. For example, in the case of DENV, the dataset was grouped according the four genetically distinct serotypes, and then further sub-grouped according to the 10 proteins coded by the proteome. For viruses with multiple groups, a combined dataset is also necessary for the purpose of combined analysis. For example, a dataset containing the NS1 proteins of the four dengue virus serotypes.

5.2.2 Identification of conserved sequences

The identification of conserved sequence fragments is an initial step to cover pathogen genomic variation. In some cases, such as HIV, influenza A viruses and dengue viruses, this variation is extensive. Multiple sequence alignments of pathogen proteins are examined by a consensus-sequence based approach (Novitsky *et al.*, 2002) for the selection of sequences conserved in the large majority of variants. For pathogens with multiple groups (clades, serotypes or subtypes), pan-group consensus sequences should be obtained by aligning consensus sequences derived from each of the different groups (Figure 5.1), rather than by analyzing pan-group alignments that combine sequences from all groups. This prevents over-represented groups from biasing the derived consensus sequence. Identification of conserved alignment sites is based on the representation (frequency) of the consensus residue among all sequences in the alignment. Depending on the variability exhibited by different pathogen groups, the cut-off intra-group representation for conserved sequences may be set from 50% to 100%. For example, in the DV analysis, conserved sites common across the four serotypes were selected, exhibiting at least 80% representation in each of the four serotypes (Figure 5.2). For immunological applications, a minimum conserved

sequence length of nine amino acids is required because this represents the typical length of peptides that bind to HLA molecules (Rammensee *et al.*, 1995).

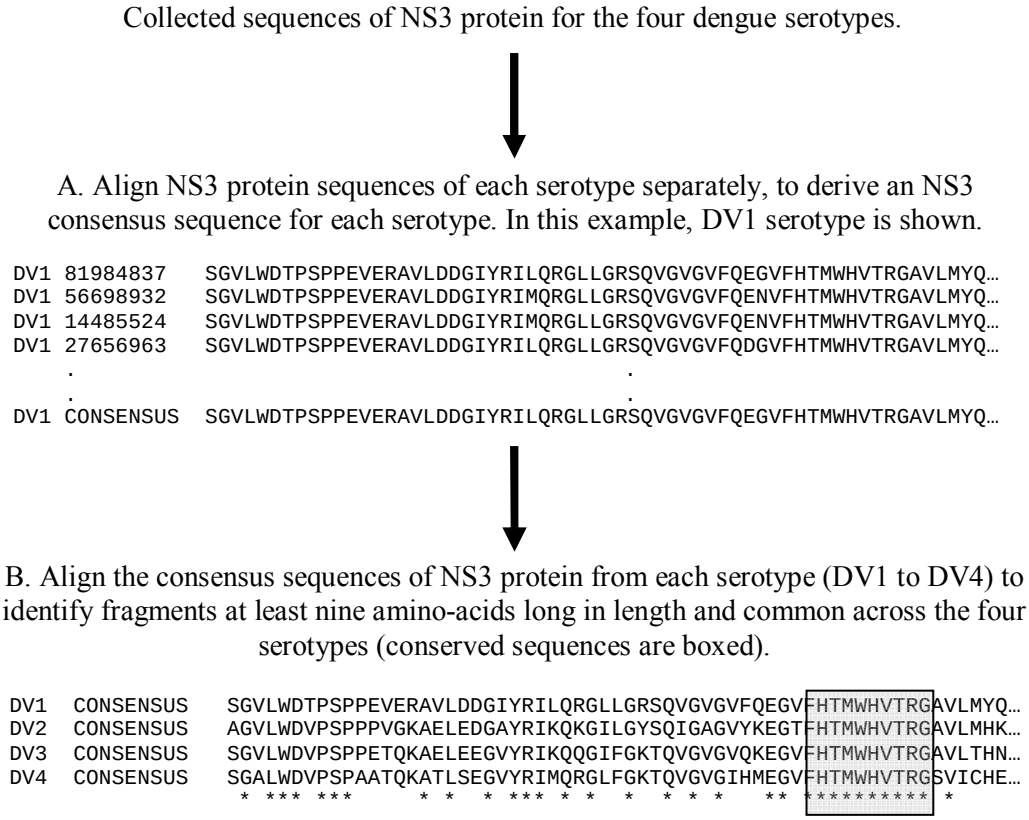


Figure 5.1: Steps involved in determining sequence fragments conserved across the four serotypes in NS3 protein using a consensus-sequence-based approach. A) The consensus sequence for NS3 protein is derived for each serotype (DV1-4) from their respective multiple sequence alignment. Each residue in the consensus sequence represents the predominant residue at that position in the corresponding multiple sequence alignment. B) The four consensus sequences of NS3 protein (one from each serotype) are aligned to reveal sequence fragments that are at least nine amino acids long and identical across the four consensus sequences.

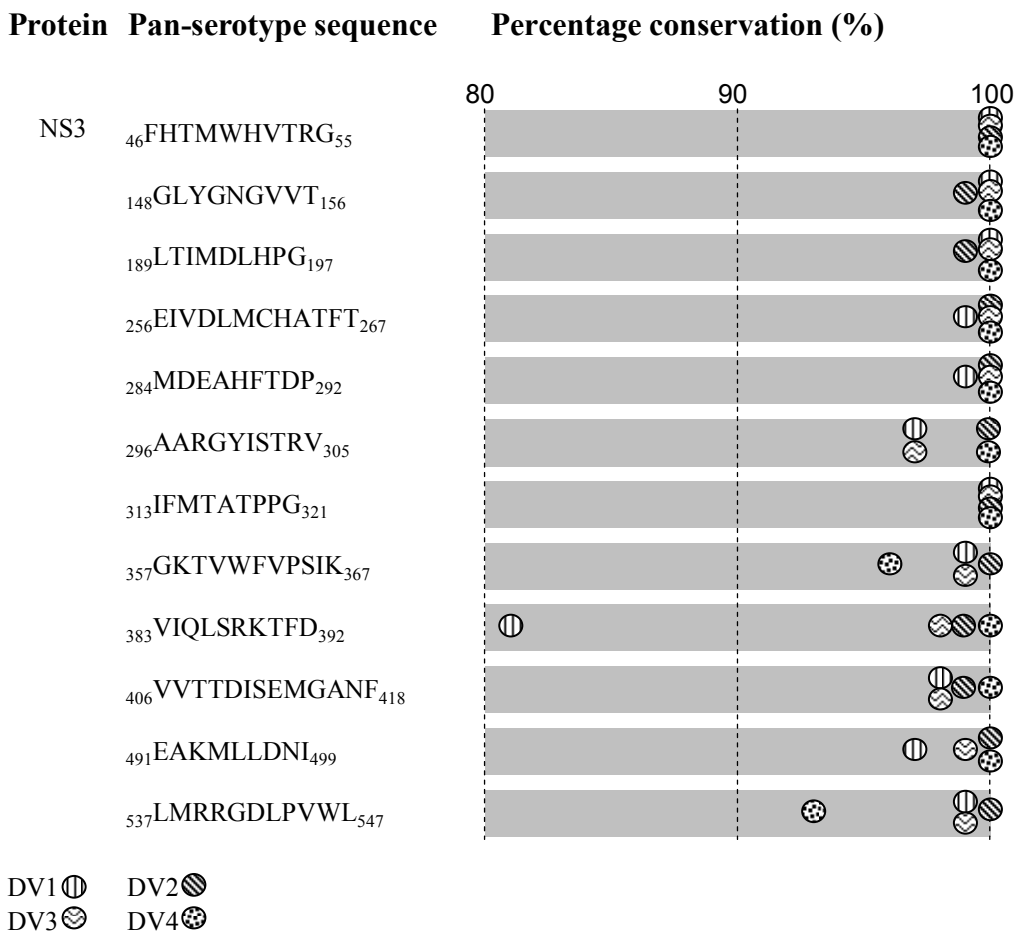


Figure 5.2: Dengue pan-serotype conserved sequences of the NS3 protein and their intra-serotype representation. The amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes.

5.2.3 Entropy-based analysis of conserved sequence variability

Consensus-based methods consider each alignment site independently. However, vaccine targets are short peptides, typically 9-mers, whose combinatorial composition can produce great diversity even when adjacent sites have highly conserved residues. A more robust method based on entropy (Shannon, 1948) can measure the degree of variability of peptides of any length, and infer their evolutionary stability. Entropy, H, representing the variability of nonamer peptides (*9*-mers) centered at any given alignment site, is computed from the probability, p_a of each nonamer peptide a occurring at that site:

$$H = -\sum_{a} p_a \log_2(p_a)$$

Peptides centered at any given position partially overlap peptides centered at neighbouring positions. Low entropy characterizes stable peptides, and an entropy value of 0 indicates a 100% conserved nonamer. Entropy rises with increasing variability of a site, and is affected both by the number of peptides at that site, and also by their respective frequency. The AVANA antigenic variability analyzer tool (Miotto *et al.*, 2008) can perform peptide entropy analysis. Figure 5.3 shows intra- and pan-serotype peptide entropy plots for dengue virus NS3 protein. The data shows that each of the four serotypes has distinct patterns of highly conserved and variable regions. Thus, the pan-serotype low entropy regions are restricted to discrete short regions, which correspond to the conserved sequences selected by the consensus-sequence method.

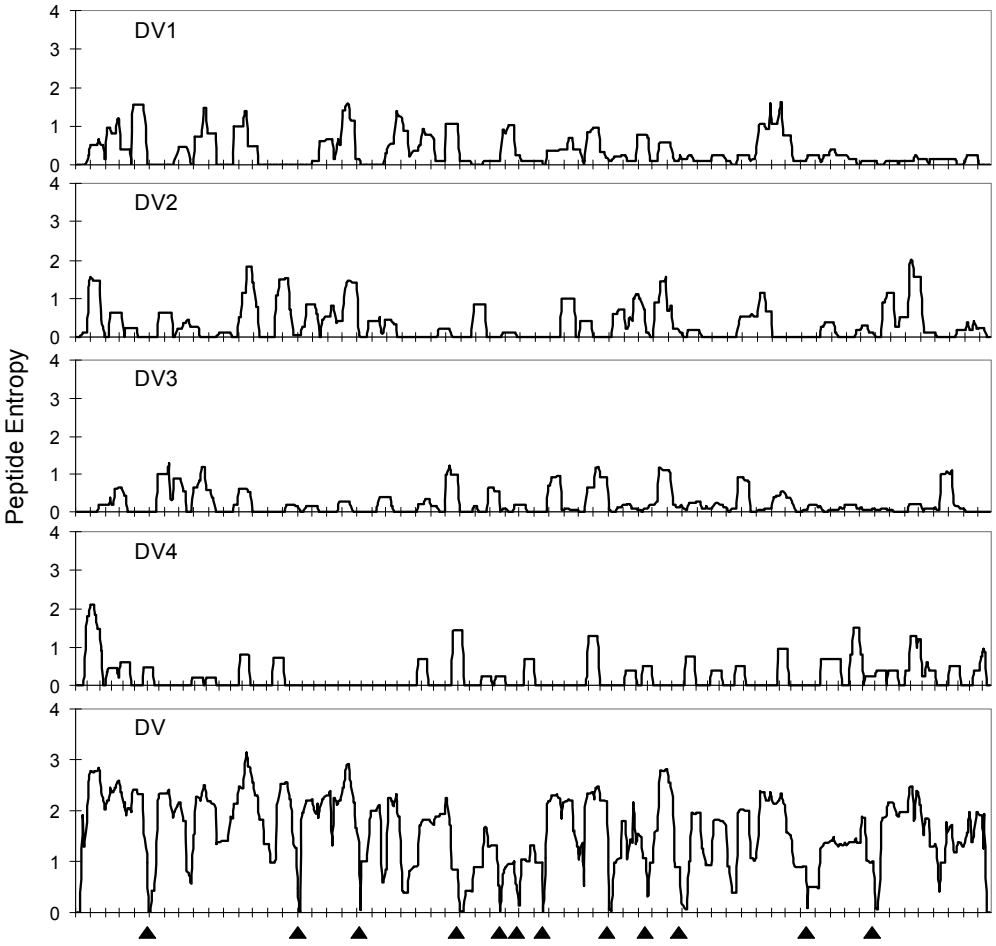


Figure 5.3: Peptide entropy plots for intra- and pan-serotype alignments of dengue virus NS3 protein (intra-serotype: DV1, DV2, DV3, DV4; pan-serotype: DV). The peptide entropy value at each position is based on the frequency of nonamer peptides present at that position in the protein's alignment. All 12 identified pan-serotype conserved sequences of NS3 protein were found to be localized in the pan-serotype conserved antigenic regions of the protein (▲), with values ranging from 0 to 0.4, indicating the high probability that these sequences will remain conserved in the future.

5.2.4 Functional and structural correlates of conserved sequences

It is recognized that conserved protein sequences generally represent important functional domains (Valdar, 2002), and their mutations would be detrimental to the survival of the pathogen. The functions of conserved sequences can be elucidated by databases that comprise data on protein families, domains and functional sites, such as the Pfam (Bateman *et al.*, 2004) (www.sanger.ac.uk/Software/Pfam) and Prosite

(Hulo *et al.*, 2006) (au.expasy.org/prosite) databases. Mapping the location of a conserved sequence on the 3-D structure of the protein may also provide relevant information (Figure 5.4). Many such 3-D structures are available in the PDB (Berman *et al.*, 2000).

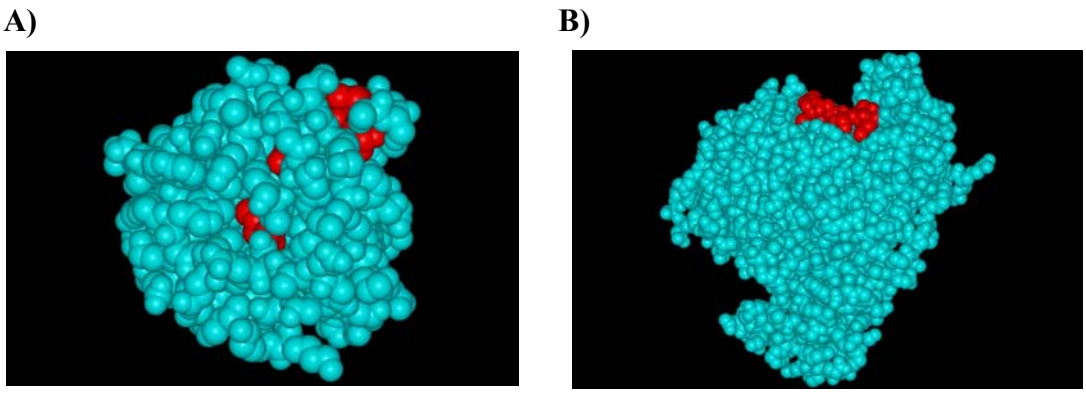


Figure 5.4: Molecular location of dengue NS3 pan-serotype conserved sequences ($_{148}$GLYGNGVVT$_{156}$ and $_{189}$LTIMDLHPG$_{197}$) shown on the 3-D structure. A) A major portion of $_{148}$GLYGNGVVT$_{156}$ conserved sequence (in red) is localized in the buried region of the 3-D structure. B) Most of the $_{189}$LTIMDLHPG$_{197}$ conserved sequence (in red) is localized in the exposed region of the 3-D structure. This suggests that the conserved sequence $_{148}$GLYGNGVVT$_{156}$ is less likely to mutate compared to $_{189}$LTIMDLHPG$_{197}$, though both share identical level of intra-serotype percentage representation (Haydon and Woolhouse, 1998).

5.2.5 Distribution of conserved sequences in nature

Potential vaccine targets should be analyzed for specificity to the target pathogen. In vaccine design, epitopes common to other pathogens could either be useful by inducing cross-protection, or detrimental by inducing altered peptide ligand effect (Rothman, 2004; Sloan-Lancaster and Allen, 1996; Evavold *et al.*, 1993). Identified conserved sequences should therefore be submitted to a BLAST search against all protein sequences at NCBI, excluding the target pathogen. If the sequences are found in other pathogens, the extent of their representation should be analyzed. For example,

many dengue virus conserved sequences are found widely present in other flaviviruses.

5.2.6 Characterization of candidate promiscuous T-cell epitopes

5.2.6.1 Algorithms for prediction of HLA binding peptides.

Dedicated algorithms based on distinct prediction models are used to locate putative promiscuous T-cell epitopes for HLA class I or II supertypes within conserved sequences. Computational epitope prediction systems, such as NetCTL (Larsen *et al.*, 2005) (www.cbs.dtu.dk/services/NetCTL), MULTIPRED (Zhang *et al.*, 2005b) and TEPITOPE (Bian and Hammer, 2004) have been proven to be effective in accurately mapping T-cell epitopes. When selecting peptides for experimental validation, putative epitopes predicted by multiple models are chosen, since consensus predictions from a combination of models have been shown to be more accurate than individual model predictions (Donnes and Kohlbacher, 2005; Larsen *et al.*, 2005).

In addition to being promiscuous with respect to multiple alleles of an HLA supertype, some putative T-cell epitopes exhibit multiple-supertype promiscuity. This additional form of promiscuity has been observed in several viruses, such as dengue (Gagnon *et al.*, 1996) and HIV (Wilson *et al.*, 2003). T-cell epitopes specific to multiple HLA supertypes are advantageous for vaccine design because they effectively increase the numbers of epitopes to which an individual can respond, and provide much more extensive coverage of the population (Wilson *et al.*, 2003).

5.2.6.2 Immunological hotspots.

Putative promiscuous T-cell epitopes may be localized in clusters, as reported in studies of HIV-1 (Brown *et al.*, 2003; Surman *et al.*, 2001; Shankar *et al.*, 1996; Berzofsky *et al.*, 1991) and the outer membrane of *Chlamydia trachomatis* (Kim and DeMars, 2001), among others (Gupta *et al.*, 2006; Srinivasan *et al.*, 2004). The clusters are also ideal for developing peptide-based vaccines because they contain multiple promiscuous epitopes. MULTIPRED (Zhang *et al.*, 2005b) and Hotspot Hunter (Zhang *et al.*, 2008) can be used to predict HLA supertype-specific immunological hotspots in pathogen sequences.

5.2.7 Altered ligand effects

The genotypic differences between primary and secondary pathogens, or between the vaccine and challenge infection, constitute a critical consideration for protective and, in some cases, pathologic immunity (Rothman, 2004). Because of intra- and inter-group sequence variability, most T-cell epitope sequences may contain single or multiple amino acid differences within and between the groups. Variants of the putative promiscuous T-cell epitopes are identified among the reported sequences in the pathogen groups, and their representation within the group and across groups is observed. Variants of a putative epitope at a given alignment position comprise all nonamers at that site that possess at least one amino acid difference. Putative epitopes with no or low variant representation (~100% conserved) are potentially advantageous in avoiding altered peptide ligand effects.

5.2.8 Experimental Validation

5.2.8.1 Survey of reported human T-cell epitopes within the conserved sequences

Predictions of T-cell epitopes of the conserved sequences can in many cases be conformed (commonly without identification of the specific allele, however) by reports of experimentally confirmed T-cell epitopes. Therefore, search against both extant literature and the Immune Epitope Database (www.immuneepitope.org) is performed for reported human T-cell epitopes (both class I and II) that fully or partially overlap with identified conserved sequences. For example, eight reported human NS3 T-cell epitopes of DV correspond to the predicted promiscuous T-cell epitopes in the NS3 conserved sequences (Table 5.1).

Table 5.1: Human T-cell epitopes in dengue virus NS3 pan-serotype conserved sequences.

Protein	Pan-serotype sequence [a]	Reported T-cell epitopes
		Reference(s)
NS3	$_{46}$FHTMWHVTRG$_{55}$	(Simmons *et al.*, 2005)
	$_{148}$GLYGNGVVT$_{156}$	(Simmons *et al.*, 2005; Kurane *et al.*, 1995)
	$_{189}$LTIMDLHPG$_{197}$	(Mangada and Rothman, 2005)
	$_{256}$EIVDLMCHATFT$_{267}$	(Simmons *et al.*, 2005; Okamoto *et al.*, 1998; Kurane *et al.*, 1993)
	$_{313}$IFMTATPPG$_{321}$	(Simmons *et al.*, 2005)
	$_{357}$GKTVWFVPSIK$_{367}$	(Loke *et al.*, 2001; Mathew *et al.*, 1996)
	$_{383}$VIQLSRKTFD$_{392}$	(Simmons *et al.*, 2005)
	$_{406}$VVTTDISEMGANF$_{418}$	(Simmons *et al.*, 2005)
	$_{537}$LMRRGDLPVWL$_{547}$	(Simmons *et al.*, 2005)

[a] *The amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes*

5.2.8.2 Experimental validation of bioinformatics screening

Experimental measurements for validation of computational predictions are necessary for accurate interpretation of results. Such measurements currently include HLA binding assays (Sidney *et al.*, 1998), immunization of HLA transgenic mice and ELISpot assay for peptide-specific T-cell activation (Rosloniec *et al.*, 1997) and of pathogen infected human subjects. Dr. Nascimento (Johns Hopkins University, USA), a collaborator of the author performed functional assessment of the dengue virus NS1 conserved sequences: four were predicted to contain HLA-DR epitopes and three of these four were confirmed by ELISpot assay with T-cell activation peptides that closely mimic the conserved sequences (Table 5.2). An additional two that were also ELISpot positive were not predicted to bind to DR molecules. In summary, of seven conserved NS1 sequences, five contain HLA-DR T-cell epitopes and at least three are promiscuous for multiple HLA-DR alleles. The predictive models are helpful in selecting antigen sequences for additional study of immune responses, especially for sequences predicted by multiple algorithms.

Table 5.2: IFN-gamma ELISpot responses of CD4-depleted splenocytes from HLA transgenic mice immunized with peptides overlapping dengue virus NS1 pan-serotype conserved sequences.

Pan-serotype sequence [a]	Predicted DR-2, -3, -4 [b]	ELISpot positive HLA transgenic mouse [c]	ELISpot activation peptide [d]
$_{12}$ELKCGSGIF$_{20}$	DR-2	DR-2	$_{13}$LKCGSGIFVTNEVHT$_{27}$
$_{25}$VHTWTEQYKFQ$_{35}$	DR-4	DR-3 and -4	$_{25}$VHTWTEQYKFQADSP$_{39}$
$_{193}$AVHADMGYWIES$_{204}$	DR-2 and -3	None	$_{193}$AVHADMGYWIESQKN$_{207}$
$_{229}$HTLWSNGVLES$_{239}$	DR-3 and -4	DR-3 and -4	$_{229}$HTLWSNGVLESDMII$_{243}$
$_{266}$GPWHLGKLE$_{274}$	None	DR-3 and -4	$_{265}$AGPWHLGKLELDFNY$_{279}$
$_{294}$RGPSLRTTT$_{302}$	None	DR-4	$_{293}$TRGPSLRTTTVSGKL$_{307}$
$_{325}$GEDGCWYGMEIRP$_{337}$	None	None	$_{325}$GEDGCWYGMEIRPIS$_{339}$

[a] T*he amino acid positions are numbered according to the aligned sequences of dengue proteins from all four serotypes*

[b] Pr*ediction for DR alleles was performed by use of MULTIPRED (Zhang et al., 2005b), TEPITOPE (Bian and Hammer, 2004) and ARB (Bui et al., 2005)*

[c] Th*e ELISpot assays were performed for DR-2, DR-3 and DR-4 transgenic mice*

[d] E*LISpot activation peptides are the actual peptides used to test the ELISpot*

5.3 Conclusion

The bioinformatics pipeline developed for DENV proved generic as it was successfully applied to several viruses, such as WNV (Chapter 6), a close relative of DENV, and a number of other viruses (data not shown). Thus, the approach described in this thesis represents a template for the analysis of other pathogens. It provides a novel and generalized approach to the formulation of peptide-based vaccines targeting a broad diversity of pathogens and applicable to the human population at large. This new methodology is a significant contribution to the field of reverse vaccinology (Vernikos, 2008; Ulmer *et al.*, 2006) as it enables the systematic screening and analyses of pathogen data which would otherwise be impossible to carry out experimentally, due to too many pathogen sequences (high viral diversity) and variations in immune system among individuals (extensive polymorphism of HLA). This approach therefore significantly reduces the efforts and cost of experimentation, while providing for systematic screening and analyses of pathogen proteomes.

Existing reverse vaccinology approaches focused on identifying conserved, HLA supertype restricted epitopes (Sette *et al.*, 2001; Sylvester-Hvid *et al.*, 2002; Wilson *et al.*, 2003; De Groot *et al.*, 2005) as vaccine targets, rely on application of bioinformatics/immunoinformatics tools to identify such epitopes, but with limited additional characterizations and typically do not involve study of all available pathogen data. Our approach supports analysis of all available pathogen data and

includes steps for additional characterizations of the epitopes, such as variant analysis and distribution in nature, among others. These two steps in particular have received little attention and are important to study the reported sequence population of the pathogen of interest and going beyond to those of all other species for evolutionarily related sequences that may act as variants to the conserved epitopes identified. These steps are necessary to identify conserved epitope sequences that are pathogen specific, with none or minimal number of variant sequences within or across other pathogen species. Variant epitopes are hypothesized to cause deleterious immune responses (see sections 2.2.2 and 4.4 for more information).

5.4 Chapter summary

Peptide-based vaccines provide a new strategy for prophylactic and therapeutic application of pathogen-specific immunity. A critical requirement of this strategy is the identification and selection of T-cell epitopes as vaccine targets that cover antigenic diversity. This chapter described current methodologies for the selection process, with DENV as a model system. A combination of publicly available bioinformatics algorithms and computational tools are used to screen and select potential *PE*s – conserved pathogen sequence fragments that act as potential T-cell epitopes of HLA supertype alleles. The selected sequences are tested for biological function by their activation of T-cells of HLA transgenic mice and of pathogen infected subjects. This approach provides an experimental basis for the design of pathogen specific, T-cell epitope peptide-based vaccines that are targeted to majority of the genetic variants of the pathogen, and are effective for a broad range of differences in HLAs among the global human population.

Chapter 6 Application of Antigenic Diversity Analysis Pipeline to West Nile Virus and Comparative Analysis to Dengue Virus

6.1 Introduction

WNV infects humans through incidental zoonotic transmission from birds via mosquitoes (Marfin *et al.*, 2001). The majority of infected individuals remain asymptomatic; however, about 20% experience mild flu-like symptoms, and approximately 1 in 150 develop severe illness, including meningoencephalitis (Hayes *et al.*, 2005; Petersen *et al.*, 2003). The virus is now endemic in many parts of the world, including Africa, Asia, Europe, Middle East, and most recently in the western hemisphere, including the United States, Mexico, and Canada (Lanciotti *et al.*, 1999). At present, there is no registered human vaccine or specific therapy for prevention or treatment of WNV infection.

WNV exhibits significant sequence diversity. Five distinct genotypes have been identified by phylogenetic analyses of the C-prM-E region by others (Bondre *et al*., 2007). Their complete genomes differ from each other by 20-25% (Bondre *et al.*, 2007). Herein, the author describes the identification and characterization of conserved, protein sequence fragments of WNV that contain promiscuous T-cell epitopes, and thus cover antigenic diversity (WNV *PE*s).

The bioinformatics pipeline developed for DENV proved generic and useful to other flaviviruses as it was successfully applied to several of them, such as WNV, JEV and YFV, and a number of non-flaviviruses (HV and HAV) (data not shown). The results of the analyses enable comparative analysis of *PE*s between viruses assessing their similarities and differences, which may provide insights into the design of better vaccine strategies. In this chapter, the results of the applications of the pipeline to WNV and the comparative analysis of *PE*s between WNV and DENV are described to demonstrate the usefulness of the approach to flaviviruses.

6.2 Materials and methods

6.2.1 West Nile virus (WNV) data preparation, selection and alignment

WNV is a mosquito-borne pathogen with genomic and proteomic structure similar to that of DENV (Chapter 2.1) (Horga and Fine, 2001; Petersen and Roehrig, 2001). WNV protein sequence records were retrieved from the NCBI Entrez Protein database in June 2007 by searching the NCBI taxonomy browser for WNV (taxonomy ID 11082). The sequences of the WNV proteins (C, prM, E, NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) were extracted from the downloaded database records and grouped according to the method described in Chapter 4.2.2, using sample protein sequences from the WNV reference record P06935 in the NCBI Entrez Protein database. The resulting dataset of each protein was then aligned by use of MUSCLE v3.6 (Edgar, 2004) with default parameters, followed by manual inspection and correction of misalignments. The alignments, which comprised both full length and partial sequences, were then subjected to a number of analyses. Identical sequences were not removed from the alignments, unless otherwise indicated in the sections below.

In large-scale proteomic analyses such as this study, bias may result from the collection of redundant sequences, derived from identical or highly similar WNV isolates sequenced by surveillance programs. Duplicate sequences (2,206) were retained for the analysis because they reflect the incidence of the corresponding WNV isolates in nature and, further, they do not affect the identification of WNV sequences that are completely (100%) conserved. As for the highly similar sequences, which may have been generated from large sequencing projects during single outbreaks;

their removal was deemed undesirable, since such arbitrary selection would introduce additional bias.

6.2.2 Amino acid difference between WNV protein sequences

Pairwise percentage amino acid difference for the full length unique sequences of each WNV proteins was computed by use of ClustalW 1.83 with default parameters. This was done to survey the extent of amino acid variation in the WNV data of 2007.

6.2.3 Nonamer entropy analysis of WNV sequences

Entropy analysis was carried out as described in Chapter 4.2.4 by use of AVANA, to study the evolutionary diversity of WNV protein sequences over the period which the sequences were collected.

6.2.4 Nonamer variant analysis of WNV sequences

The conservation and variability of nonamer sequences in the WNV protein alignments were further analyzed to study the representation of nonamer peptides variant to the predominant peptide at a given position *x* in the alignment, following the guidelines in the Chapter 4.2.5.

6.2.5 Identification of completely conserved WNV sequences (pan-WNV sequences)

Completely conserved sequences (pan-WNV sequences) of at least nine amino acids and fully identical in all the sequences analyzed (100% representation) were identified from the multiple sequence alignment of each protein. Peptides with the unknown

residue X in the alignments were ignored. The threshold of 100%, which is ideal, was used for this virus because of the higher level of conservation exhibited by the virus, unlike DENV.

6.2.6 Structure-function analysis of pan-WNV sequences

The reported and putative functional properties of pan-WNV sequences were searched in the literature and by use of the Prosite database and Pfam databases. The conserved sequences were also mapped onto the 3-D structures of WNV proteins if they were available in the PDB. See Chapter 4.2.6 for details.

6.2.7 Identification of pan-WNV sequences common to other viruses and organisms

Pan-WNV sequences that were common to at least nine consecutive amino acids of other viruses were identified by performing BLAST search against protein sequences of all other viruses reported at NCBI (as of August 2007) (parameters: limit by Entrez query "txid10239[Organism:exp] NOT txid11082[Organism:exp]"; "automatically adjust parameters for short sequences" option disabled; "low-complexity" filter disabled; alignments: 20,000; expect threshold: 200,000, or 20,000, or 2,000 until a valid result was obtained; word size: 2; matrix: PAM30; gap costs: "Existence: 9, Extension: 1"; compositional adjustments: no adjustment). Similar BLAST searches were carried out against protein sequences of all organisms excluding viruses (parameters: limit by Entrez query "Root[ORGN] NOT Viruses[ORGN] NOT txid81077[ORGN]"; "automatically adjust parameters for short sequences" option disabled; "low-complexity" filter disabled; alignments: 20,000; expect threshold: 200,000, or 20,000, or 2,000 until a valid result is obtained; word size: 2; matrix:

PAM30; gap costs: “Existence: 9, Extension: 1”; compositional adjustments: no adjustment). Artificial sequence hits were removed by the “NOT txid81077[ORGN]” keyword.

6.2.8 Identification of known and predicted WNV HLA-supertype binding epitopes

Both literature and the Immune Epitope Database were analyzed to identify previously reported HLA class I and II human T-cell epitopes of WNV that overlapped at least nine consecutive amino acids of pan-WNV sequences. In addition, four prediction models were used to identify candidate WNV sequences that bind to multiple HLA class I or II supertype alleles (see Chapter 4.2.8).

6.2.9 Comparative analysis of *PE*s between WNV and DENV

The identified and characterized *PE*s of WNV were compared against those of DENV, described in Chapter 4, to study the level of conservation between the viruses and identify characteristics common between the *PE*s of the viruses.

6.3 Results

6.3.1 WNV protein sequence datasets

A total of 2,746 complete and partial WNV protein sequences were retrieved from the NCBI Entrez Protein database as of June 2007 (Table 6.1). The large number of sequences analyzed and their wide spatial and temporal (1955-2007; based on WNV NCBI records with available annotation) distribution enabled a broad survey of WNV diversity in nature. The distribution of these sequences varied considerably among the

different proteins (from 141 NS4b sequences to 927 E sequences). Comparisons of amino acid variations between the full length unique sequences of the 10 WNV proteins showed that C had the highest range of amino acid differences across the sequences (up to 23%), while NS4b had the lowest (up to 8%) (Table 6.1).

6.3.2 Evolutionary stability of WNV

The evolutionary diversity of WNV was studied by computing entropy for analysis of immunologically relevant nonamer peptide sequences. The entropy plot revealed the evolutionary variability of nonamer sequences across the WNV proteome (Figure 6.1). The vast majority of nonamer positions exhibited low to moderate entropy (≤ 1.0), indicating lower probability of mutations occurring over time. Many regions had zero entropy signifying no change throughout the recorded history of the virus. Peak or near peak entropy values (~2) were observed in the E, NS4a and NS5 proteins. The NS5 protein known to be one of the most conserved across *Flavivirus* proteins, had the highest percentage of completely conserved nonamer regions, but also exhibited high entropy in regions at the C-terminal of the protein. Overall, entropy analysis revealed numerous highly conserved and evolutionarily stable WNV sequences distributed throughout the viral proteins, indicative of high genetic stability of WNV, despite its adaptability to global emergence.

Table 6.1: Number of WNV protein sequences retrieved from NCBI and their maximum percentage amino acid difference over the protein length.

WNV protein	Total length (aa) [b]	No. of sequences analysed [a]	% maximum amino-acid difference [c]
C	123	264	23
prM	167	417	19
E	497	927	12
NS1	352	164	16
NS2a	231	143	20
NS2b	131	146	10
NS3	619	146	10
NS4a	149	142	14
NS4b	256	141	8
NS5	905	256	10
Total	*3430*	*2746*	-

[a] *Retrieved from NCBI Entrez Protein database on 28th June 2007*

[b] *Approximate size indicated in number of amino acids*

[c] *Maximum percentage amino-acid difference for each WNV protein, computed using ClustalW*

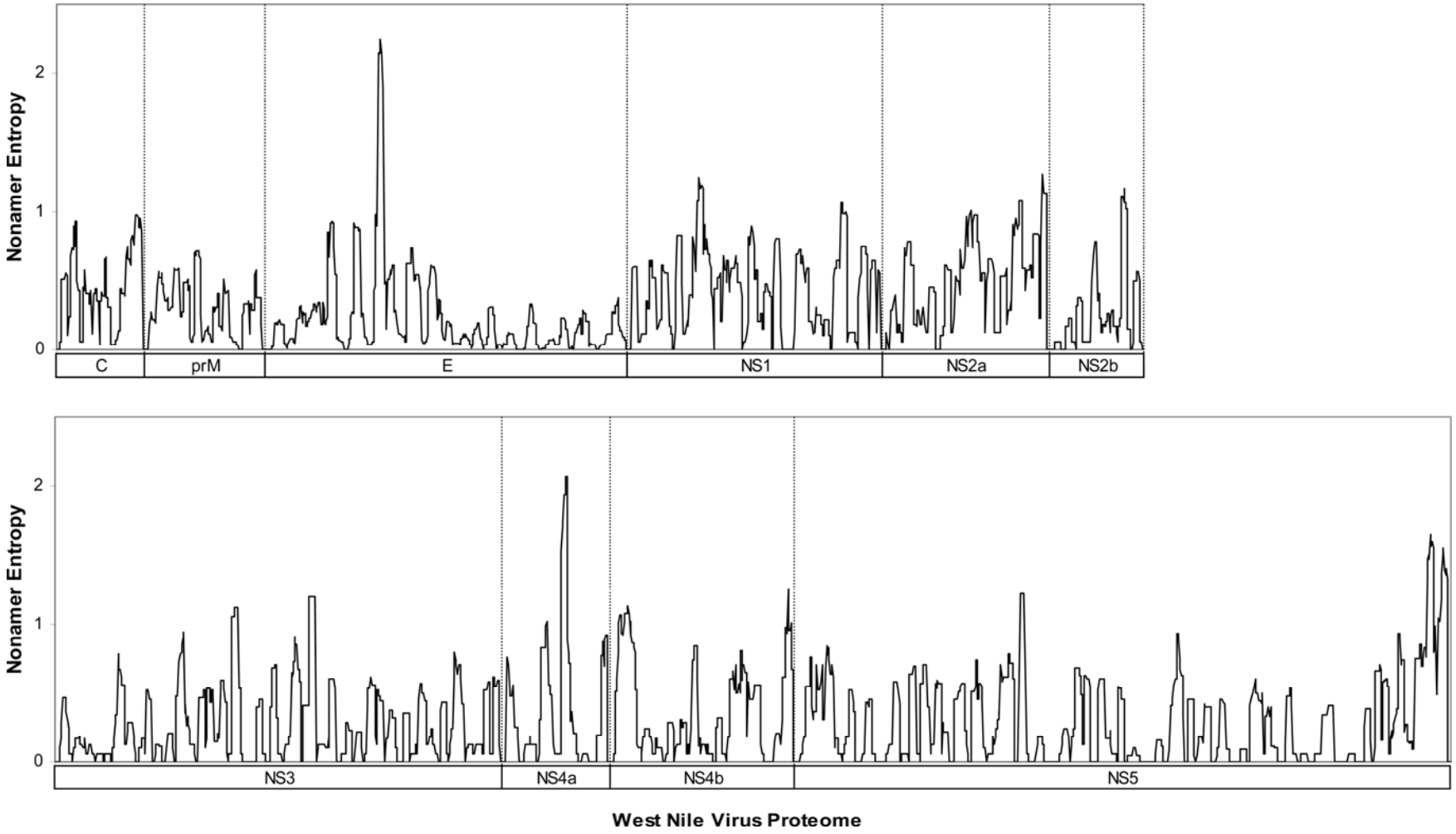


Figure 6.1: Peptide entropy plots for WNV protein alignments.

6.3.3 Representation of variant WNV sequences

Completely conserved nonamer site with zero variants were numerous and the occurrence of variant nonamer sequences across the WNV proteome was generally low, less than 10% of all WNV recorded sequences at most positions (Figure 6.2). The position with the highest representation of variant nonamer sequences (49%) was found in the nonstructural protein NS4a. Overall, the data suggests a low probability of immune response challenges from variant WNV T-cell epitopes, due to a high representation of historically conserved sequences of the WNV proteome in the known virus population.

6.3.4 Completely conserved pan-WNV sequences

A total of 88 completely conserved sequence fragments (pan-WNV sequences) were identified across the whole proteome (Table 6.2). The length of these fragments ranged from 9 to 29 amino acids, covering a total length of 1,169 amino acids (~34%) of the complete WNV polyprotein length (3,430 aa) (Table 6.3). The C protein had no pan-WNV sequence, which is consistent with the large number of amino acid differences (23%) observed in this protein compared to other WNV proteins (Table 6.1). The NS3 and NS5 proteins contained the greatest number of pan-WNV sequences, 25 in NS3 (spanning 48% of the protein length) and 30 in NS5 (51% of the protein length). The other nonstructural proteins NS1, NS2a, NS2b, NS4a and NS4b collectively contained a total of 24 completely conserved sequences, covering 11% to 40% of their respective protein lengths. In contrast, the variability of the structural proteins was much greater: prM had only two pan-WNV sequences (14% of the protein length), while E had seven (18% of the protein length).

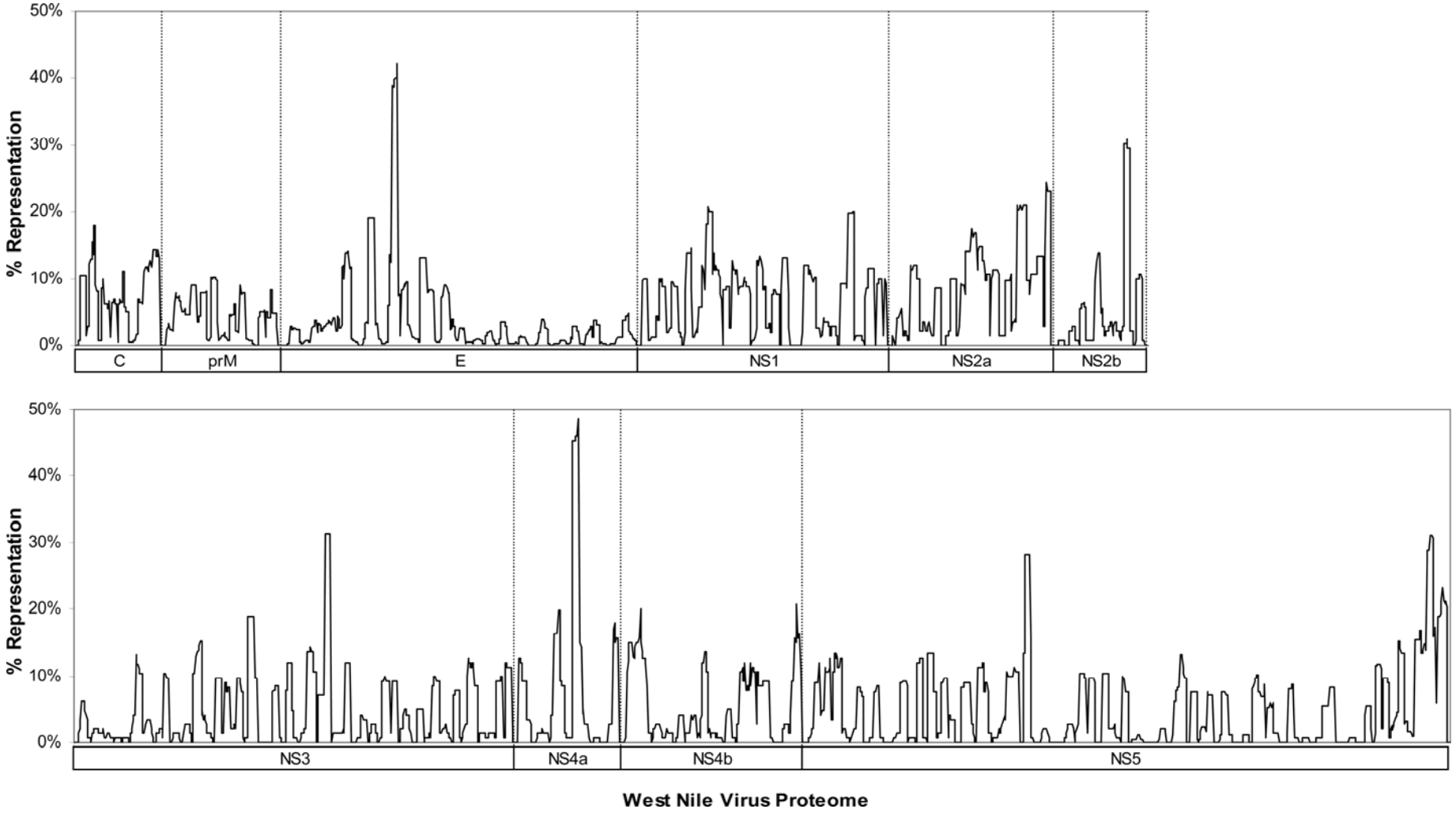


Figure 6.2: Percentage representation of nonamer variants in relation to the predominant nonamer peptide for all nonamer positions in WNV protein alignments.

Table 6.2: Completely conserved sequence fragments (pan-WNV sequences) of WNV proteins.

WNV protein	Length (aa)	Pan-WNV sequence [a]
C	-	None
prM	14	125-ESWILRNPGYALVA-138
	10	158-LLLLVAPAYS-167
E	11	001-FNCLGMSNRDF-011
	14	104-GCGLFGKGSIDTCA-117
	9	293-LKGTTYGVC-301
	19	338-SVASLNDLTPVGRLVTVNP-356
	12	370-ELEPPFGDSYIV-381
	10	417-LGDTAWDFGS-426
	12	449-LFGGMSWITQGL-460
NS1	11	058-RSVSRLEHQMW-068
	9	114-GWKAWGKSI-122
	9	154-EVEDFGFGL-162
	10	195-HSDLSYWIES-204
	25	209-TWKLERAVLGEVKSCTWPETHTLWG-233
	11	276-DFDYCPGTTVT-286
	10	313-CRSCTLPPLR-322
	10	328-GCWYGMEIRP-337
NS2a	10	004-DMIDPFQLGL-013
	15	69-NSGGDVVHLALMATF-83
NS2b	10	001-GWPATEVMTA-010
	14	012-GLMFAIVGGLAELD-025
	9	032-PMTIAGLMF-040
	11	108-SAYTPWAILPS-118
NS3	10	001-GGVLWDTPSP-010
	10	020-TGVYRIMTRG-029
	10	052-TTKGAALMSG-061
	10	063-GRLDPYWGSV-072
	10	074-EDRLCYGGPW-083
	12	108-NVQTKPGVFKTP-119
	12	131-PTGTSGSPIVDK-142
	13	145-DVIGLYGNGVIMP-157
	11	161-YISAIVQGERM-171
	12	191-VLDLHPGAGKTR-202
	9	235-ALRGLPIRY-243
	27	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282
	14	288-HFTDPASIAARGYI-301
	12	310-AAAIFMTATPPG-321
	10	337-QTEIPDRAWN-346
	10	357-GKTVWFVPSV-366
	16	385-QLNRKSYETEYPKCKN-400
	11	408-TTDISEMGANF-418
	11	422-RVIDSRKSVKP-432
	11	451-TAASAAQRRGR-461
	16	470-GDEYCYGGHTNEDDSN-485
	9	487-AHWTEARIM-495
	11	526-LRGEERKNFLE-536
	9	540-TADLPVWLA-548
	10	563-WCFDGPRTNT-572
NS4a	15	019-KTWEALDTMYVVATA-033

WNV protein	Length (aa)	Pan-WNV sequence [a]
	11	043-ALEELPDALQT-053
	13	101-GTKIAGMLLLSLL-113
	20	115-MIVLIPEPEKQRSQTDNQLA-134
NS4b	9	039-PATAWSLYA-047
	13	068-TSLTSINVQASAL-080
	9	085-RGFPFVDVG-093
	12	138-AQRRTAAGIMKN-149
	10	156-VATDVPELER-165
	22	208-VTLWENGASSVWNATTAIGLCH-229
NS5	10	001-GGAKGRTLGE-010
	9	060-AKLRWLVER-068
	17	079-DLGCGRGGWCYYMATQK-095
	22	107-GPGHEEPQLVQSYGWNIVTMKS-128
	10	141-DTLLCDIGES-150
	10	152-SSAEVEEHRT-161
	9	168-VEDWLHRGP-176
	16	208-RNPLSRNSTHEMYWVS-223
	12	235-MTSQVLLGRMEK-246
	10	259-NLGSGTRAVG-268
	13	299-NHPYRTWNYHGSY-311
	18	318-SASSLVNGVVRLLSKPWD-335
	29	340-VTTMAMTDTTPFGQQRVFKEKVDTKAPEP-368
	10	375-VLNETTNWLW-384
	18	404-KVNSNAALGAMFEEQNQW-421
	10	440-EREAHLRGEC-449
	12	451-TCIYNMMGKREK-462
	29	472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500
	16	504-NSGGGVEGLGLQKLGY-519
	13	533-YADDTAGWDTRIT-545
	10	548-DLENEAKVLE-557
	15	571-IELTYRHKVVKVMRP-585
	23	596-ISREDQRGSGQVVTYALNTFTNL-618
	12	620-VQLVRMMEGEGV-631
	19	662-RMAVSGDDCVVKPLDDRFA-680
	14	689-MSKVRKDIQEWKPS-702
	18	704-GWYDWQQVPFCSNHFTEL-721
	27	741-GRARISPGAGWNVRDTACLAKSYAQMW-767
	21	769-LLYFHRRDLRLMANAICSAVP-789
	12	792-WVPTGRTTWSIH-803

[a] *Numbers prefixing and affixing sequences represent start and end positions in the protein alignment*

Table 6.3: Number of pan-WNV sequences, their length in amino acids and percentage coverage of total protein length.

WNV protein	Total length (aa) [a]	Pan-WNV sequences		
		Number	Length (aa)	% of total protein length (aa) [b]
C	123	0	0	0
prM	167	2	24	14
E	497	7	87	18
NS1	352	8	95	27
NS2a	231	2	25	11
NS2b	131	4	44	34
NS3	619	25	296	48
NS4a	149	4	59	40
NS4b	256	6	75	29
NS5	905	30	464	51
Total	3430	88	1169	34

[a] *Approximate length indicated in number of amino acids, according to the reference protein sequence described in the Methods*

[b] *Approximate percentage rounded off to nearest whole number*

6.3.5 Functional and structural analysis of pan-WNV sequences

Sequences conserved throughout the evolutionary history of rapidly mutating RNA viruses are thought to be critical for structure and/or function. A search in the Prosite and Pfam databases revealed that 50 of the 88 pan-WNV sequences are known to be associated with putative or known biological functions and/or structure (Table 6.4); the biological significance of the remaining 38 sequences is still to be determined. In the E protein, two pan-WNV sequences correspond to the fusion loop and dimerisation domain (Kanai *et al.*, 2006), while two correspond to immunoglobulin-like domain, attributed to putative receptor binding sites (Mukhopadhyay *et al.*, 2003). One NS1 sequence corresponds to the putative ATP/GTP binding site p-loop motif, likely to be involved in helicase activity (Li *et al.*, 1999). NS3 contained four

pan-WNV sequences that correspond to the peptidase family S7 (*Flavivirus* serine protease) domain (Erbel *et al.*, 2006), and four that correspond to known/putative *Flavivirus* Asp-Glu-Ala-Asp/His (DEAD/H) domain associated with ATP-dependent helicase activity (Feito *et al.*, 2008). NS5 contained 17 sequences corresponding to the RNA dependent RNA polymerase (RdRp)/catalytic domain (Mackenzie *et al.*, 2007; Malet *et al.*, 2007). Furthermore, 33 of the 50 pan-WNV sequences were predicted to exhibit post-translational modification(s), including N-glycosylation, protein kinase C (PKC), casein kinase II (CKII) and tyrosine kinase (TK) phosphorylation, N-myristoylation and/or amidation.

Amino acid residues exposed and protruding on the surface of viral proteins are generally subject to fewer packing constraints and residue interactions as compared to those buried within protein interiors. Thirty of the 88 pan-WNV sequences could be mapped on available, but incomplete, WNV protein structures obtained from the PDB (E protein, 2HG0; NS3, 2IJO; and NS5, 2HFZ) (Appendix 6). Five pan-WNV sequences were mostly buried and an equal number of pan-WNV sequences were partially exposed (13) or largely exposed (12). These results should be considered preliminary until full length 3-D structures are available.

Table 6.4: Reported biological properties of pan-WNV sequences.

WNV protein	Pan-WNV sequence	Functional domains and motifs [a]	Putative post-transcriptional modifications [a]
E	1-FNCLGMSNRDF-11	Dimerisation domain	PKC, CKII
	104-GCGLFGKGSIDTCA-117	Dimerisation domain, Fusion Loop	N-myristoylation
	293-LKGTTYGVC-301	-	N-myristoylation
	338-SVASLNDLTPVGRLVTVNP-356	Immunoglobulin-like domain	CKII
	370-ELEPPFGDSYIV-381	Immunoglobulin-like domain	-
	417-LGDTAWDFGS-426	-	CKII
NS1	58-RSVSRLEHQMW-68	-	CKII
	114-GWKAWGKSI-122	ATP/GTP-binding site motif A (P-loop)	-
	209-TWKLERAVLGEVKSCTWPETHTLWG-233	-	PKC, CKII
	328-GCWYGMEIRP-337	-	N-myristoylation
NS2a	69-NSGGDVVHLALMATF-83	-	CKII
NS2b	12-GLMFAIVGGLAELD-25	-	N-myristoylation
NS3	52-TTKGAALMSG-61	-	PKC
	74-EDRLCYGGPW-83	Peptidase S7	-
	108-NVQTKPGVFKTP-119	Peptidase S7	-
	131-PTGTSGSPIVDK-142	Peptidase S7	-
	145-DVIGLYGNGVIMP-157	Peptidase S7	N-myristoylation
	191-VLDLHPGAGKTR-202	DEAD/H domain	N-myristoylation
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282	DEAD/H domain	PKC
	288-HFTDPASIAARGYI-301	DEAD/H domain	-
	310-AAAIFMTATPPG-321	DEAD/H domain	-
	385-QLNRKSYETEYPKCKN-400	-	TK

	422-RVIDSRKSVKP-432	-	PKC
	470-GDEYCYGGHTNEDDSN-485	-	CKII, N-myristoylation
NS4a	101-GTKIAGMLLLSLL-113	-	N-myristoylation
	115-MIVLIPEPEKQRSQTDNQLA-134	-	CKII
NS4b	208-VTLWENGASSVWNATTAIGLCH-229	-	N-glycosylation, CKII
NS5	1-GGAKGRTLGE-10	-	CKII, N-myristoylation
	79-DLGCGRGGWCYYMATQK-95	-	PKC, N-myristoylation
	107-GPGHEEPQLVQSYGWNIVTMKS-128	-	PKC
	152-SSAEVEEHRT-161	-	CKII
	208-RNPLSRNSTHEMYWVS-223	-	N-glycosylation, CKII
	299-NHPYRTWNYHGSY-311	RdRp	-
	318-SASSLVNGVVRLLSKPWD-335	RdRp	-
	340-VTTMAMTDTTPFGQQRVFKEKVDTKAPEP-368	RdRp	-
	375-VLNETTNWLW-384	-	N-glycosylation
	404-KVNSNAALGAMFEEQNQW-421	RdRp	-
	451-TCIYNMMGKREK-462	RdRp	Amidation
	472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500	RdRp	-
	504-NSGGGVEGLGLQKLGY-519	RdRp	N-myristoylation
	533-YADDTAGWDTRIT-545	RdRp	-
	571-IELTYRHKVVKVMRP-585	RdRp	PKC
	596-ISREDQRGSGQVVTYALNTFTNL-618	RdRp/ RdRp catalytic domain	CKII, N-myristoylation
	620-VQLVRMMEGEGV-631	RdRp	-
	662-RMAVSGDDCVVKPLDDRFA-680	RdRp/ RdRp catalytic domain	CKII
	689-MSKVRKDIQEWKPS-702	RdRp	-
	704-GWYDWQQVPFCSNHFTEL-721	RdRp	-
	741-GRARISPGAGWNVRDTACLAKSYAQMW-767	RdRp	N-myristoylation
	769-LLYFHRRDLRLMANAICSAVP-789	RdRp	-

792-WVPTGRTTWSIH-803	RdRp	PKC

[a] *Prosite (PS) and Pfam (PF) accession numbers: PS00001, N-glycosylation site; PS00005, Protein kinase C phosphorylation (PKC) site; PS00006, Casein kinase II (CKII) phosphorylation site; PS00007, tyrosine kinase (TK) phosphorylation site; PS00008, N-myristoylation site; PS00009, Amidation site; PS00017, ATP/GTP-binding site motif A (P-loop); PS50507, RNA-directed RNA polymerase (RdRp) catalytic domain; PF00869, dimerisation domain; PF00949, Peptidase S7; PF00972, RNA-directed RNA polymerase (RdRp); PF02832, Immunoglobulin-like domain; PF07652, Flavivirus DEAD/H domain.*

6.3.6 Distribution of pan-WNV sequences in nature

Sixty-seven (67) of the 88 pan-WNV sequences (~76%) overlapped at least nine amino acid sequences of as many as 68 other viruses of the family *Flaviviridae*, genus *Flavivirus* (Figure 6.3). Each of these 67 sequences matched at least one and at most 61 *Flavivirus* species (Figure 6.4 and Appendix 7). Murray valley encephalitis virus shared 49 of the 67 pan-WNV sequences; Japanese encephalitis and Usutu virus shared 47 and 41, respectively; and representatives of some of the important human pathogens, St. Louis encephalitis, dengue, tick-borne encephalitis, and yellow fever viruses shared from 36 to 11 of the 67 pan-WNV sequences. The representation of these pan-WNV sequences ranged from low to high across reported sequences of the several well studied flaviviruses, including dengue (DENV), Japanese encephalitis (JEV), Louping ill (LIV), Omsk hemorrhagic fever (OMSK), Powassan (PV), St. Louis encephalitis (LEV), Tick-borne encephalitis (TBEV), and Yellow fever (YFV) (Appendix 7). For example, the pan-WNV sequence $E_{104\text{-}117}$ was present in 99% of the 245 E protein JEV sequences, while $E_{293\text{-}301}$ was present in only 1% of the 256 E protein JEV sequences.

Fifty-eight (58) of the 67 pan-WNV sequences shared by other flaviviruses were from the non-structural proteins. Of the 27 pan-WNV sequences found in NS5, 10 were present in at least 30 *Flavivirus* species; while of the 16 represented in NS3, three were found in between 25 and 34 other species; the remaining 15 sequences were contained in non-structural proteins NS1 (7), NS2a (2), NS2b (1), NS4a (3) and NS4b (2). Nine (9) of the 67 pan-WNV sequences shared by flaviviruses originated from the structural proteins E (7) and prM (2); one of the E protein sequences was present in 31 species.

Remarkably, five of the 88 pan-WNV sequences ($prM_{158-167}$, $NS3_{408-418}$, $NS4b_{208-229}$, $NS5_{1-10}$, and $NS5_{504-519}$) shared nine consecutive amino acids with seven non-viral species. The nonamer sequence from $prM_{158-166}$ is found in the bacterium *Acidiphilium cryptum* JF-5; $NS3_{409-417}$ in the mosquito *Aedes albopictus*; $NS4b_{218-226}$ in the Japanese rice *Oryza sativa* (japonica cultivar-group); $NS5_{2-10}$ in the bacterium *Actinomyces odontolyticus*; $NS5_{504-512}$ in the bacteria *Burkholderia ambifaria* MC40-6 and *Burkholderia cepacia* AMMD; and $NS_{506-514}$ in the bacterium *Methylobacterium extorquens* PA1.

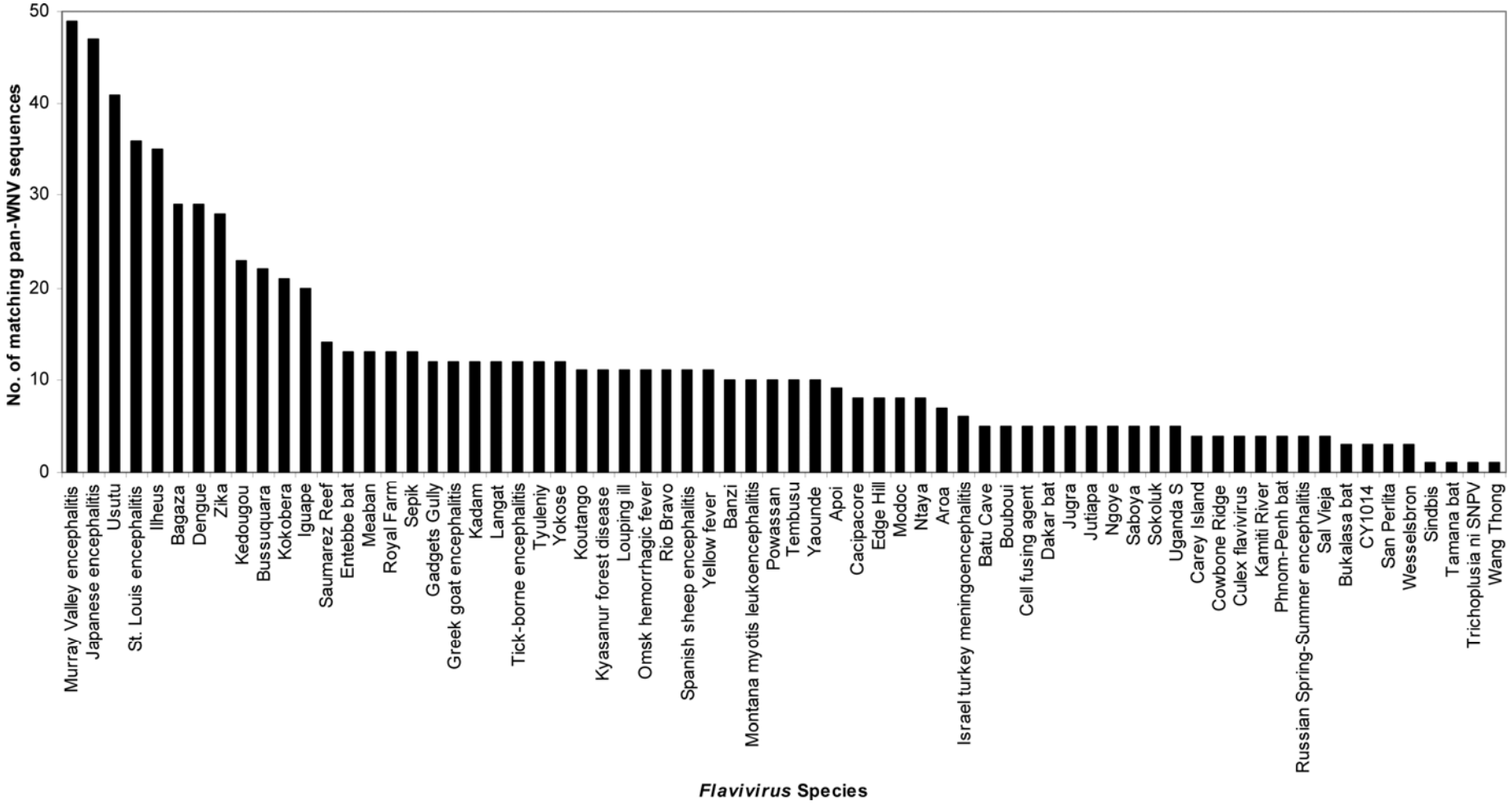


Figure 6.3: Pan-WNV sequences conserved in other flaviviruses.

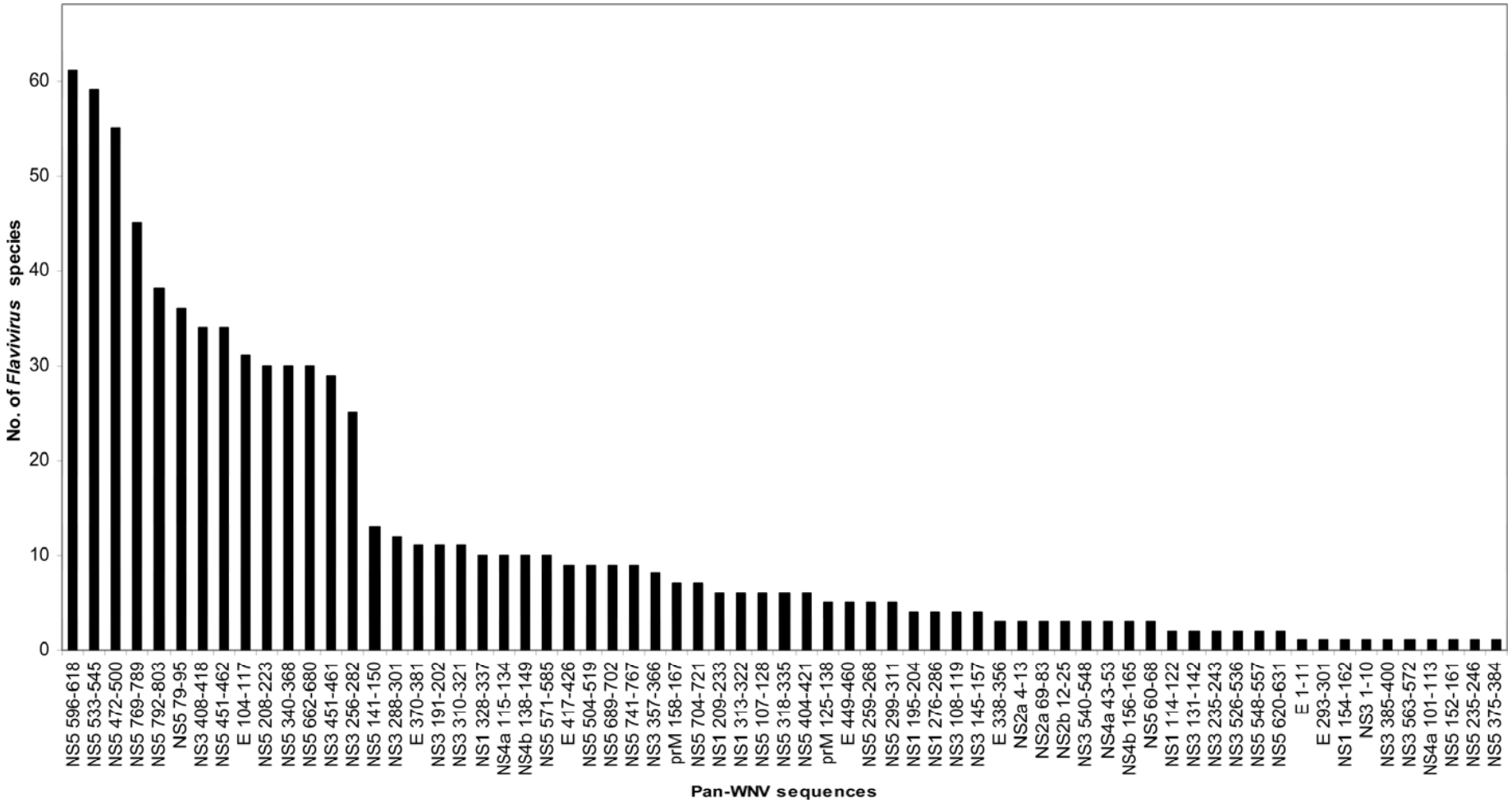


Figure 6.4: Number of other flaviviruses sharing the pan-WNV sequences.

6.3.7 Known and predicted HLA supertype-restricted, pan-WNV T-cell epitopes

Literature survey and IEDB database search revealed that three of the pan-WNV sequences (2 in NS3, and one in NS5) overlapped at least nine amino acids of three previously reported WNV T-cell epitopes immunogenic in human, with their HLA restriction, when known, showed both class I (B*07) and II (DR2) specificities (Table 6.5). Further evaluation of the immune-relevance of pan-WNV sequences included a search for candidate putative promiscuous HLA supertype-restricted T-cell epitopes within these regions by use of NetCTL, Multipred, ARB and TEPITOPE prediction tools. Seventy-eight (78) of the 88 pan-WNV sequences (~89%) (Figure 6.5) was predicted to contain 271 supertype-restricted binding nonamers (Appendix 8). Of these sequences, 62 contain nonamers predicted to bind to multiple HLA supertypes. Clusters of predicted binders, two or more overlapping nonamer peptides, with identical HLA supertype-restrictions, known as hotspots (Zhang *et al.*, 2008; Zhang *et al.*, 2005b), were found in 41 of the 78 sequences. Seven (7) of the 78 sequences had three sequential nonamers overlapping by eight amino acids. As these sequences are completely conserved, all of these epitopes are found in all reported WNV strains.

In addition, 44 pan-WNV sequences were found to contain sequences of at least nine amino acids present in 54 $CD4^{+}CD8^{-}$ and/or $CD4^{-}CD8^{+}$ IFN-γ ELISpot positive peptides (Table 6.6), identified by peptide-specific T-cell responses of murine H-2 class I or II-deficient transgenic mice, expressing prototypic class I HLA-A2 (A*0201), -A24 (A*2402) and -B7 (B*0702), and class II HLA-DR2 (DRB1*1501), -DR3 (DRB1*0301) and -DR4 (DRB1*0401) alleles, and immunized with overlapping peptides covering the entire WNV proteome (unpublished data of our collaborator Jung KO *et al*., of Johns Hopkins University, Maryland, USA). Twenty-three (23) of

the 44 pan-WNV sequences that overlapped the ELISpot positive peptides correspond to positive HLA-DR supertype-restricted T-cell epitope predictions by either Multipred or TEPITOPE (Table 6.6 and Appendix 8). The experimental data revealed that 11 out of 44 pan-WNV sequences, localized in prM, E, NS1, NS3, NS4a, NS4b and NS5, were promiscuous for at least two HLA-DR alleles; the promiscuity of nine of these 11 pan-WNV sequences were correctly predicted (Appendix 8). In summary, combined with previously reported data for human WNV T-cell epitopes from literature and public database (Table 6.5), at least 44 of the 88 pan-WNV sequences contain numerous HLA-restricted class I and/or class II epitopes demonstrated by *in vivo* T-cell response assays.

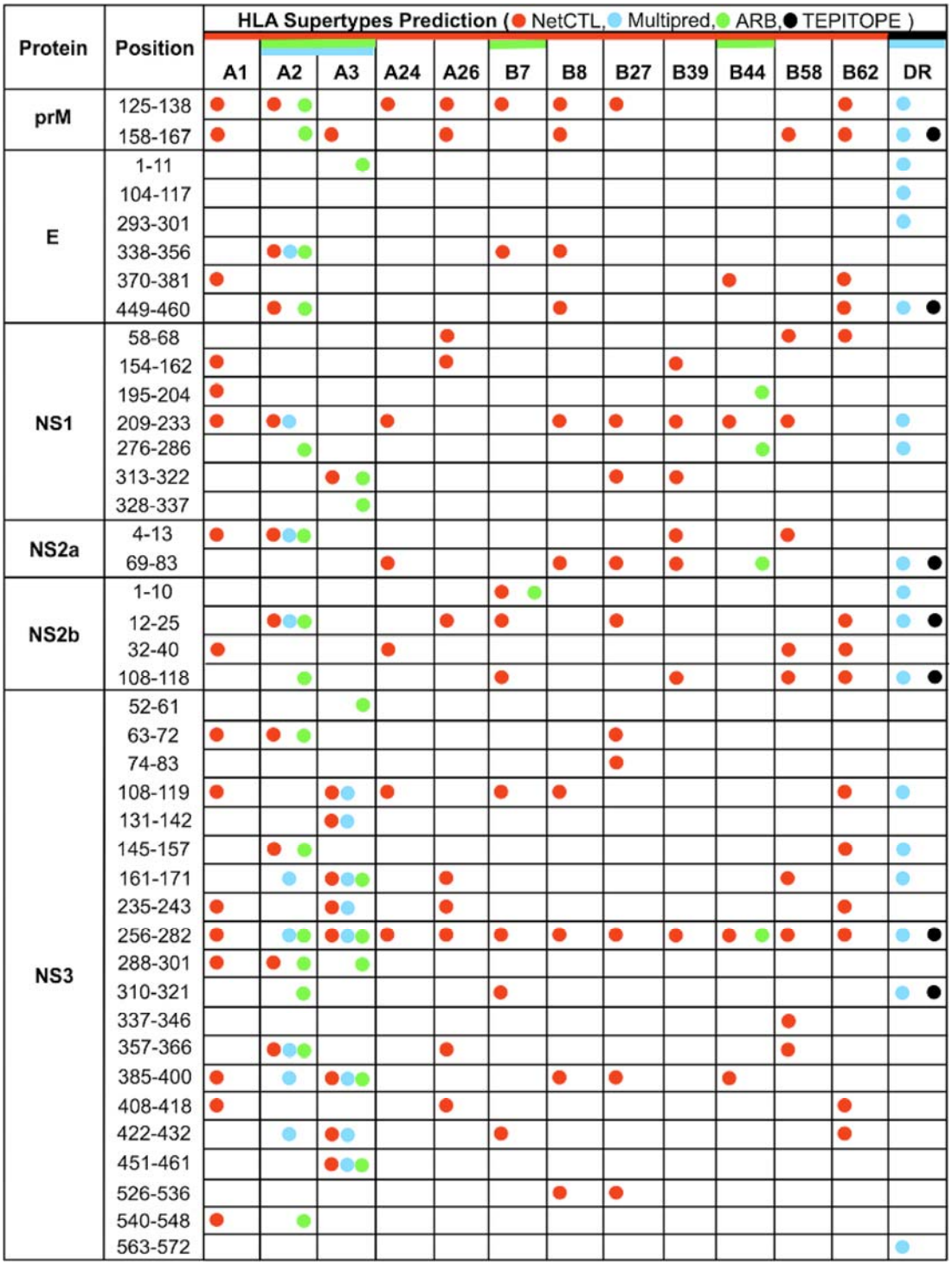
Table 6.5: WNV sequences with human T-cell epitopes elucidated by other studies.

WNV protein	Pan-WNV sequence	Reported T-cell epitopes immunogenic in humans			
		Sequence [a]	T-cells	HLA restriction	Reference(s) [b]
NS3	145-DVIGLYGNGVIMP-157	VIGLYGNGV	CD4	DR2	(Kurane *et al.*, 1995)
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282	SPHRVPNYNL	CD8	B07	(De Groot *et al.*, 2001)
NS5	704-GWYDWQQVPFCSNHFTEL-721	FCSNHFTEL	-	-	1021472

[a] *Epitope amino acids matching the pan-WNV sequences are underlined*

[b] *1021472 is an accession number of a record in the Immune Epitope Database*

A


HLA Supertypes Prediction (NetCTL, Multipred, ARB, TEPITOPE)
Protein
Position
A1
A2
A3
A24
A26
B7
B8
B27
B39
B44
B58
B62
DR
prM
125-138
158-167
E
1-11
104-117
293-301
338-356
370-381
449-460
NS1
58-68
154-162
195-204
209-233
276-286
313-322
328-337
NS2a
4-13
69-83
NS2b
1-10
12-25
32-40
108-118
NS3
52-61
63-72
74-83
108-119
131-142
145-157
161-171
235-243
256-282
288-301
310-321
337-346
357-366
385-400
408-418
422-432
451-461
526-536
540-548
563-572

B

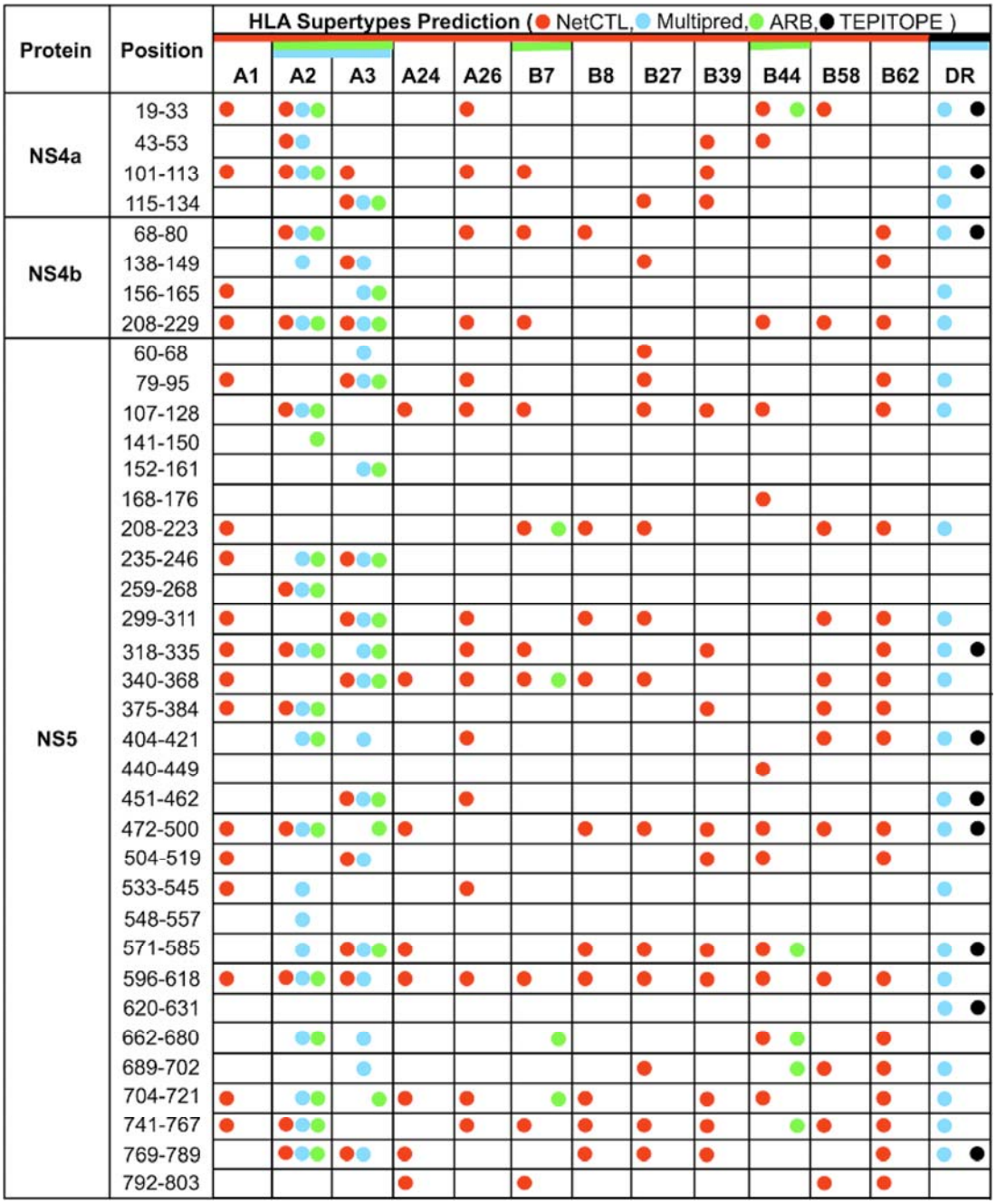


Figure 6.5: Candidate HLA supertype-restricted, pan-WNV T-cell epitopes predicted by computational algorithms. Results for C, prM, E, NS1, NS2a, NS2b and NS3 protein are shown in panel A, while panel B for NS4a, NS4b, and NS5.

Table 6.6: Pan-WNV sequences with human T-cell epitopes identified by use of HLA transgenic mice.

WNV protein	Pan-WNV sequence [a]	T-cell epitopes immunogenic in HLA transgenic mice	
		ELISpot activation peptide [b]	ELISpot positive HLA transgenic mouse
prM	125-ESWILRNPGYALVA-138*	LVKTESWILRNPGYALVA	DR2 & DR4
		LRNPGYALVAAVIGWML	A24, B7 & DR2
	158-LLLLVAPAYS-167*	RVVFVVLLLLVAPAYS	A2, DR2, DR3 & DR4
E	104-GCGLFGKGSIDTCA-117*	RGWGNGCGLFGKGSI	DR3 & DR4
	293-LKGTTYGVC-301*	EKLQLKGTTYGVCSKAFK	DR4
	370-ELEPPFGDSYIV-381	KVLIELEPPFGDSYIVV	DR4
	449-LFGGMSWITQGL-460*	FRSLFGGMSWITQGLLGA	A2, DR2 & DR3
NS1	209-TWKLERAVLGEVKSCTWPETHTLWG-233*	RLNDTWKLERAVLGEVK	DR4
	276-DFDYCPGTTVT-286*	EGRVEIDFDYCPGTTVTL	DR4
	313-CRSCTLPPLR-322	GKLITDWCCRSCTLPPLR	DR3 & DR4
	328-GCWYGMEIRP-337	SGCWYGMEIRPQRHDEK	DR4
NS2a	69-NSGGDVVHLALMATF-83*	FAESNSGGDVVHLALMA	DR4
NS2b	1-GWPATEVMTA-10*	GWPATEVMTAVGLMFAIV	DR4
	108-SAYTPWAILPS-118*	ISAYTPWAILPSVVGFWI	A24, B7 & DR4
NS3	1-GGVLWDTPSP-10	GGVLWDTPSPKEYKK	B7 & DR4
	52-TTKGAALMSG-61	WHTTKGAALMSGEGRL	DR3
	074-EDRLCYGGPW-083	GSVKEDRLCYGGPWKLQH	A2
	145-DVIGLYGNGVIMP-157	PIVDKNGDVIGLYGNGVI	A2
		VIGLYGNGVIMPNGSYI	A2

	161-YISAIVQGERM-171*	YISAIVQGERMDEPIPA	A2 & DR2
	191-VLDLHPGAGKTR-202	MLRKKQITVLDLHPGAGK	A2 & DR2
		VLDLHPGAGKTRRILPQI	DR2
	235-ALRGLPIRY-243	VAAEMAEALRGLPIRY	DR4
		EALRGLPIRYQTSAVPR	DR4
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282*	PREHNGNEIVDVMCHATL	A2, DR2 & DR4
		IVDVMCHATLTHRLMSPH	DR2
		TLTHRLMSPHRVPNYNLF	A2 & DR2
	310-AAAIFMTATPPG-321	KVELGEAAAIFMTATPPG	A2
	337-QTEIPDRAWN-346	LQTEIPDRAWNSGYEWI	A2
	422-RVIDSRKSVKP-432	EMGANFKASRVIDSRKSV	A2
	470-GDEYCYGGHTNEDDSN-485	CYGGHTNEDDSNFAHW	A2 & DR3
	487-AHWTEARIM-495	AHWTEARIMLDNINM	A2 & DR3
	526-LRGEERKNFLE-536	EYRLRGEERKNFLELLR	A2 & DR2
	563-WCFDGPRTNT-572*	DRRWCFDGPRTNTIL	DR3
NS4a	19-KTWEALDTMYVVATA-33*	HFMGKTWEALDTMYVVA	DR2 & DR4
	115-MIVLIPEPEKQRSQTDNQLA-134*	VLIPEPEKQRSQTDNQLA	DR4
NS4b	39-PATAWSLYA-47	GEFLLDLRPATAWSLYAV	DR2
		PATAWSLYAVTTAVLTPL	DR2 & DR3
	68-TSLTSINVQASAL-80*	DYINTSLTSINVQASALF	DR3 & DR4
	208-VTLWENGASSVWNATTAIGLCH-229*	LITAAAVTLWENGASSVW	DR3 & DR4
NS5	107-GPGHEEPQLVQSYGWNIVTMKS-128*	LVQSYGWNIVTMKSGVDV	DR3
	152-SSAEVEEHRT-161	CDIGESSSSAEVEEHRTI	B7
		SAEVEEHRTIRVLEMV	A2, B7 & DR2

208-RNPLSRNSTHEMYWVS-223*	SRNSTHEMYWVSRASGNV	DR2
451-TCIYNMMGKREK-462	ECHTCIYNMMGKREKK	A2
472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500	AKGSRAIWFMWLGARFL	A24
	WFMWLGARFLEFEALGFL	A24
596-ISREDQRGSGQVVTYALNTFTNL-618*	REDQRGSGQVVTYALNTF	DR2
	GQVVTYALNTFTNLAVQL	DR2 & DR4
620-VQLVRMMEGEGV-631*	NTFTNLAVQLVRMMEGEGV	DR4
704-GWYDWQQVPFCSNHFTEL-721*	GWYDWQQVPFCSNHFTEL	DR4
741-GRARISPGAGWNVRDTACLAKSYAQMW-767	DTACLAKSYAQMWLLLYF	A24
769-LLYFHRRDLRLMANAICSAVP-789*	YAQMWLLLYFHRRDLRLM	B7 & DR4
792-WVPTGRTTWSIH-803	NWVPTGRTTWSIHAGGEW	DR4

[a] *Pan-WNV sequences that are predicted, either by Multipred or TEPITOPE, to contain at least one HLA-DR supertype-restricted binding nonamer are indicated by an asterisk (*)*

[b] *Epitope amino acids matching the pan-WNV sequences are underlined*

6.3.8 Similarities and differences between *PE*s of WNV and DENV

In contrast to WNV (average entropy: 0.23 to 0.51), the sequences of the combined serotypes of DENV are highly diverse (average inter-serotype entropy: 1.6 to 2.6), with only 44 pan-DENV sequences, representing 15% of the proteome length, that are present in 80% or more of the sequences of each serotype; unlike WNV (88 pan-WNV sequences of 100% representation representing 34% of the proteome), only two of the 44 sequences were completely conserved in all the four serotypes (2007 dataset). However, the conservation and variability of the individual DENV serotypes (average intra-serotype entropy: 0.2 to 1.0) is comparable to WNV.

Despite the differences in the number, length and representation of the pan-DENV and pan-WNV sequences, they share several common characteristics. Both pan-DENV and pan-WNV sequences have shown remarkable stability over the entire recorded history of their sequences. It is likely that these sequences have been under selection pressure to fulfill critical biological and/or structural properties (a number of them have been shown to be important for structure and function). In addition, majority of these sequences of both viruses shared extensive conservation with other flaviviruses (76% of pan-WNV sequences matched 68 flaviviruses, while 61% of pan-DENV sequences matched 64), in particular those of DENV, despite the great variability of the virus. Consequently, the number of species specific sequences for both viruses was similar (17 pan-DENV and 21 pan-WNV sequences). Interestingly, a number of both pan-WNV and pan-DENV sequences displayed conservation extending to non-viruses, suggesting active recombination between phyla. In addition, many of the pan-DENV and pan-WNV sequences are immunologicall relevant. Overall, these results insinuate that the *PE*s of other flaviviruses are also likely to share similar characteristics.

6.4 Discussion

In the 70 years following the discovery of WNV in Africa in 1937 (Smithburn *et al.*, 1940), there has been 100% conservation of 88 pan-WNV sequences in the reported data, corresponding collectively to 1169 aa or ~34% of the 3,430 aa total composition of the viral proteome. The remaining 66% of the proteome contained one or more amino acid variants within each nonamer segment across the reported WNV sequences. Most of the pan-WNV sequences were found in the non-structural proteins. Quantitatively, 40% (1058/2643 aa) of the amino acids of the non-structural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) comprised the pan-WNV sequences, compared to only 14% (111/787 aa) of the structural proteins (C, prM and E). This marked difference in the evolutionary conservation/variability of the viral proteins can be attributed to greater demands on the integrity of nonstructural proteins in their viral functional roles (Lindenbach and Rice, 2003), and possibly to the selective advantage of modified structural proteins in the adaptation to host immune responses. This evolutionary history of the conserved protein sequences extends to other members of the *Flaviviridae* family, with 67 of the 88 pan-WNV sequences shared among at least 68 other flaviviruses. Many of the identified critical biological and/or structural properties are associated with the conserved sequences; for example, the E dimerisation domain and fusion loop (Kanai *et al.*, 2006; Mukhopadhyay *et al.*, 2003), NS3 peptidase S7, DEAD/H domain (Feito *et al.*, 2008; Erbel *et al.*, 2006), and NS5 proteins RdRp domain (Mackenzie *et al.*, 2007; Malet *et al.*, 2007). Hence, these conserved sequences are unlikely to significantly diverge in newly emerging WNV isolates in the future, and represent attractive targets for the development of diagnostics, specific anti-viral compounds and vaccine candidate targets. In short, they can be defined as multi-purpose immutable, functional and immunological tags

of WNV.

It is also noteworthy that nine consecutive amino acids of five of the pan-WNV sequences are also present in non-viral proteomes, the *Aedes albopictus* mosquito, *Oryza sativa* Japanese rice and several bacteria. This overlap of pan-WNV sequences with non-viral sequences is possibly coincidental, but is likely to be statistically significant as the probability of randomly matching a nonamer is almost negligible ($1/(20^9)$). WNV protein sequences found in the proteomes of bacteria are possibly due to integration of some unknown virus into the bacterial genome (Biswas *et al.*, 2005; Gottesman and Weisberg, 2004). Similarly, the NS3 nonamer sequence fragment found in the Asian Tiger mosquito (*Aedes albopictus*), is possibly due to genetic recombination between phyla (Crochu *et al.*, 2004). Unexpectedly, a nonamer of WNV NS4b protein was found in a single instance within a plant pathogenesis-related protein from Japanese rice (*Oryza sativa*), which functions as plant defense system against pathogens (Freeman, 2003).

There is evidence that many of the conserved sequences are immunologically relevant in humans. Numerous (44/88) contain at least nine amino acids overlapping with a total of 54 peptides that have been reported to be immunogenic in humans and/or HLA transgenic mice. In addition, putative T-cell epitopes were predicted by computational analysis for 12 major HLA class I supertypes and for class II DR supertype, with broad application to the immune responses of human population worldwide. Some of the putative T-cell epitopes were predicted to be promiscuous to multiple HLA supertypes as has been observed with several viruses (Khan *et al.*, 2008). These findings of the limited variability of WNV sequences relevant to cellular immunity point to the probable success in the development of a WNV vaccine as compared to the history of failure of candidate vaccines against the much more highly

variable *Flavivirus*, such as DENV (Khan *et al.*, 2008).

The results obtained herein enabled a comparative analysis of *PE*s between DENV and WNV by comparing and contrasting the number of identified *PE*s and their characteristics, such as i) future conservation potential (entropy analysis), ii) conservation breadth - number of other viruses sharing the exact sequence of the *PE* of the target pathogen (*i.e.* at least nine consecutive amino acids threshold used herein), iii) conservation depth - frequency or representation of the *PE* sequence of the target pathogen in all known sequences of the other viruses that share the sequence, iv) functional-structural relevance, v) altered peptide ligand potential by variants of the target pathogen or other viruses that share the *PE* and vi) immunogenicity potential. Such comparative analysis help i) quantify the level of conservation of the viruses being compared, and ii) identify features common or different to *PE*s of viruses of interest, which will contribute to better understanding of *PE*s across pathogens and may provide insights into better design of vaccine strategies.

6.5 Chapter summary

***Background*:** The systematic bioinformatics approach described in Chapter 5 was applied to WNV, a close relative of DENV, to demonstrate the generic nature of the approach and perform a comparative analysis to DENV of sequences that cover antigenic diversity. The comparative analysis will help elucidate similarities and differences in the characteristics of *PE*s between pathogens of interest, which may provide insights into the design of better vaccine strategies.

***Results*:** The author describes a large-scale analysis of the entire WNV proteome, aimed at identifying and characterizing WNV *PE*s. This study, which used

2,746 WNV protein sequences collected from the NCBI Entrez Protein database, focused on analysis of peptides of nine amino acids or more, which are immunologically relevant as potential T-cell epitopes. Entropy-based analysis of the diversity of WNV sequences, revealed the presence of numerous evolutionarily stable nonamer positions across the proteome (entropy value of ≤ 1). The representation (frequency) of nonamers variant to the predominant peptide at these stable positions was, generally, low (≤ 10% of the WNV sequences analyzed). Eighty-eight fragments of length 9-29 amino acids, representing ~34% of the WNV polyprotein length, were identified to be identical and evolutionarily stable in all analyzed WNV sequences. Of the 88 completely conserved sequences, 67 are also present in other flaviviruses, and several have been associated with the functional and structural properties of viral proteins. Immunoinformatic analysis revealed that the majority (78/88) of conserved sequences are potentially immunogenic, while 44 contain experimentally confirmed human T-cell epitopes. The results obtained herein enabled a comparative analysis of *PE*s between DENV and WNV. In contrast to WNV, the sequences of the combined serotypes of DENV are highly diverse, with only 44 pan-DENV sequences; however, the conservation and variability of the individual DENV serotypes is comparable to WNV. Further, despite the differences in the number, length and representation of the DENV and WNV *PE*s, they share several common characteristics.

***Conclusions*:** This study identified a comprehensive catalogue of completely conserved WNV sequences, many of which are shared by other flaviviruses, and a majority is potential epitope. These sequences constitute as potential WNV candidate *PE*s. The complete conservation of these immunologically relevant sequences through the entire recorded WNV history suggests they will be valuable as components of peptide-specific vaccines or other therapeutic applications, for sequence-specific

diagnosis of a wide-range of *Flavivirus* infections, and for studies of homologous sequences among other flaviviruses. The identification and characterization of WNV *PE*s enabled a comparative analysis of *PE*s between DENV and WNV, which helped quantify the level of conservation of the viruses being compared and identify common characteristics of the *PE*s between the viruses.

Chapter 7 Conservation Patterns of *PE*s across Dengue Virus and Other Members of the Genus *Flavivirus*

7.1 Introduction

The family *Flaviviridae* consists of evolutionary related flaviviruses that have common ancestral origin, genomic architecture and their proteins have similar functional/structural role (Solomon and Mallewa, 2001; Henchal and Putnak, 1990). Our antigenic diversity analysis of both DENV and WNV (described in Chapter 4 and 6) revealed extensive conservation of the majority of the *PE*s of each virus across numerous flaviviruses. This similarity could possibly have positive (cross-protection) or negative (altered ligand effect), or both implications to vaccine design, in part depending on the two dimensions of conservation – the breadth and depth. We define breadth as the number of viruses that share the *PE* sequence of interest, while depth is the level of representation of the *PE* in each of the viruses shared. For example, the *PE* $_{141}$DTLLCDIGESS$_{151}$ (pan-DENV sequence of the NS5 protein) was observed to be common across 13 different flaviviruses (breadth), with a depth ranging from low (< 1%), high (84%) to complete representation (100%) in the reported sequences of these viruses (Table 4.7). *PE*s with extensive breadth and high depth are good candidates for pan-*Flavivirus* peptide-based vaccine design.

However, because the analysis of depth and breadth of *PE*s only considered the flaviviruses that shared the specific *PE*, it does not provide a holistic view of the conservation pattern of a given *PE* to evolutionary related flaviviruses that share or do not share this *PE*. Complementing the analysis of depth and breadth, the evolutionary relationship of *PE*s across flaviviruses can be studied to better understand their conservation pattern and assess the possibility of a pan-*Flavivirus* vaccine.

In this study, the author utilized evolutionary distance to investigate conservation pattern of *PE*s across flaviviruses. In addition, the stability of the relationship observed at the *PE* level was benchmarked against at the proteome and

protein level. Phylogenetic approach is well-suited for this analysis as it provides a holistic view of the relationship amongst the sequences of the species analysed, without restricting to the perspective of a single species sequence, such as to DENV and WNV in the analysis of breadth and depth in the earlier chapters (Chapters 4 and 6). This analysis helps reveal the conservation pattern between *PE*s of viruses of the same genus in terms of evolutionary distance and may provide new insights into vaccine design. Flaviviruses are a good model for this comparative analysis because they have similar genomic architecture and code for the same 10 proteins (Henchal and Putnak, 1990).

7.2 Materials and methods

7.2.1 Data

Evolutionary analysis of DENV and 28 other flaviviruses was performed at the proteome, protein and *PE* levels using phylogenetic approach. At the time of analysis (November 2008), complete proteome (polyprotein) sequences of only 29 flaviviruses (including DENV) were available at the NCBI Entrez Protein database, which included: TBEV, Tick-borne encephalitis virus; WNV, West Nile virus; DENV, dengue virus (comprising DENV1, 2, 3, and 4); OMSK, Omsk hemorrhagic fever virus; KFDV, Kyasanur forest disease virus; KRV, Kamiti River virus; JEV, Japanese encephalitis virus; LV, Langat virus; CF, Culex flavivirus; TBV, Tamana bat virus; UV, Usutu virus; PV, Powassan virus; AP, Apoi virus; RBV, Rio Bravo virus; MVE, Murray Valley encephalitis virus; LIV, Louping ill virus; CFAV, Cell fusing agent virus; EV, Entebbe bat virus; RFV, Royal Farm virus; LEV, St. Louis encephalitis virus; MMLV, Montana myotis leukoencephalitis virus; MV, Modoc virus; YFV,

Yellow Fever virus; KBV, Kokobera virus; IH, Ilheus virus; IV, Iguape virus; BV, Bussuquara virus; SV, Sepik virus; YV, Yokose virus. These flaviviruses were considered to be representative of the genus flavivirus as they represented 11 out of the 14 *Flavivirus* taxonomy group classification available at the NCBI (Table 7.1).

Table 7.1: NCBI taxonomy group classification of selected flaviviruses [a].

Flavivirus group	*Flavivirus* group member virus species
Aroa virus	Iguape virus, Bussuquara virus
Dengue virus group	Dengue virus
Japanese encephalitis virus group	West Nile virus, Japanese encephalitis virus, Usutu virus, Murray Valley encephalitis, St. Louis encephalitis virus
Kokobera virus group	Kokobera virus
Modoc virus group	Modoc virus
Mosquito-borne viruses	Ilheus virus, Sepik virus
Ntaya virus group	Yokose virus
Rio Bravo virus group	Apoi virus, Rio Bravo virus, Entebbe bat virus
Seaborne tick-borne virus group	-
Spondweni virus group	-
Tick-borne encephalitis virus group	Tick-borne encephalitis virus, Omsk hemorrhagic fever virus, Kyasanur forest disease virus, Langat virus, Powassan virus, Louping ill virus, Royal Farm virus
Yaounde virus	-
Yellow fever virus group	Yellow Fever virus
Unclassified Flavivirus	Kamiti River virus, Culex flavivirus, Tamana bat virus, Cell fusing agent virus, Montana myotis leukoencephalitis

[a] *Data obtained from www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=11051.*

For the analysis at the protein level, all the proteins that contained the pan-DENV sequences were studied (structural: E; non-structural: NS1, NS3, NS4a, NS4b and NS5) to ensure results representative of the different proteins in the proteome. For the analysis at the *PE* level, peptides from the respective protein of the flaviviruses that corresponded to the pan-DENV sequences were extracted for the study.

7.2.2 Analysis

Phylogenetic trees were generated for the polyprotein proteomes, proteins and the *PE*s using the PROTDIST, followed by the NEIGHBOR program of PHYLIP 3.6a2 (Felsenstein, 1989). The parameters used for the two programs are as follows: PROTDIST – "JTT" for "categories model", "No" for "gamma distribution of rates among positions", "Yes" for "one category of substitution rates", "No" for "use weights for positions", and "No" for "analyze multiple data sets" option; NEIGHBOR – "neighbor-joining" for "neighbor-joining or UPGMA tree", "Yes" for "outgroup root", "No" for "lower-triangular data matrix", "No" for upper-triangular matrix, "No" for "subreplicates", "No" for "randomize input order of species", and "No" for "analyze multiple data sets" option). Hepatitis C virus genotype 6 proteome sequence (Accession no. YP_001469634.1) was used as the outgroup and rooted tree diagrams were generated with the TREEVIEW program (Page, 2002). The pan-DENV sequences $_{296}$AARGYISTRV$_{305}$ (NS3) and $_{35}$PASAWTLYAVATT$_{47}$ (NS4b) were excluded from the analysis because of technical difficulty in generating their trees using the alignment data extracted from the flaviviruses. In addition, the pan-DENV sequence $_{383}$VIQLSRKTFD$_{392}$ (NS3) was ignored because the corresponding sequence in the DENV1 strain was a rare variant.

7.3 Results

The family *Flaviviridae* consists of closely related flaviviruses. Based on the phylogenetic tree of the proteome sequences (Figure 7.1) of the selected 29 flaviviruses (including DENV), they grouped into eight clusters. The clustering patterns at the proteome and the protein level were generally consistent, particularly

evident in the E protein (Figure 7.2 A). However, minor exceptions were revealed by the analysis at the protein level (Figure 7.2 and Appendix 9). For example, the cluster at the proteome level (Figure 7.1) containing TBV, CF, KRV, and CFAV was separated into independent clusters i) CF, KRV and CFAV, and ii) TBV in the NS3 protein (Figure 7.2 B).

At the *PE* level (Figure 7.3 and Appendix 9), the clustering patterns of the flaviviruses were significantly different from the patterns observed at the proteome and protein levels. For example, the virus KBV sequence at the proteome level (Figure 7.1) formed a cluster by itself, which was close to the DENV cluster, while at the protein level (Figure 7.2) it was in the same cluster as DENV, however, at the *PE* level (pan-DENV sequence $_{97}$VDRGWGNGCGLFGKG$_{111}$ in the E protein; Figure 7.3) it not only grouped but shared the same exact sequence with viruses that were distant to it at the proteome and protein level (such as AP, EV, MMLV, RBV, SV, YFV, and YV); in contrast, at this *PE* level, DENV was distant to the KBV group. Further, the relationship observed amongst the viruses at the *PE* level varied from one *PE* to another; for example, the grouping of KBV for the *PE* $_{252}$VLGSQEGAMH$_{261}$ in the E protein was different from that of $_{97}$VDRGWGNGCGLFGKG$_{111}$ described earlier.

In addition, as observed in Chapter 4 and 6 for DENV and WNV, respectively, generally, a number of the viruses showed zero antigenic distance (shared identical sequence) at the *PE* level; however, the grouping and the number of viruses exhibiting such zero antigenic distance varied from *PE* to *PE*. For example, the identity and the number of viruses exhibiting zero antigenic diversity for *PE* $_{97}$VDRGWGNGCGLFGKG$_{111}$ in the E protein (group 1: KFDV, LIV, LV, OMSK, RBV and TBEV; group 2: AP, EV, MMLV, RBV, SV, YFV, YV, and KBV; group 3:

JEV, MVE, and UV; group 4: DENV1, DENV2, DENV3, DENV4, IH, LEV, and WNV; group 5: CF and CFAV) varied significantly from the *PE* $_{252}$VLGSQEGAMH$_{261}$ of the same protein (group 1: DENV1, DENV2, DENV3, and DENV4; group 2: BV, LEV, MVE, UV, WNV, and IV; group 3: LIV, LV, OMSK, and TBEV).

In summary, the clustering patterns at the proteome and the protein level were generally consistent, with minor exceptions for some of the proteins. At the *PE* level, the clustering patterns of the flaviviruses were significantly different from the patterns observed at the proteome and protein level. Further, the patterns varied from one *PE* to another. In addition, generally for each *PE*, a number of the viruses showed zero antigenic distance (shared identical sequence); however, the grouping and the number of viruses exhibiting such zero antigenic distance also varied from *PE* to *PE*. The main differences between the proteome, protein, and *PE* groupings are shown in Figure 7.4.

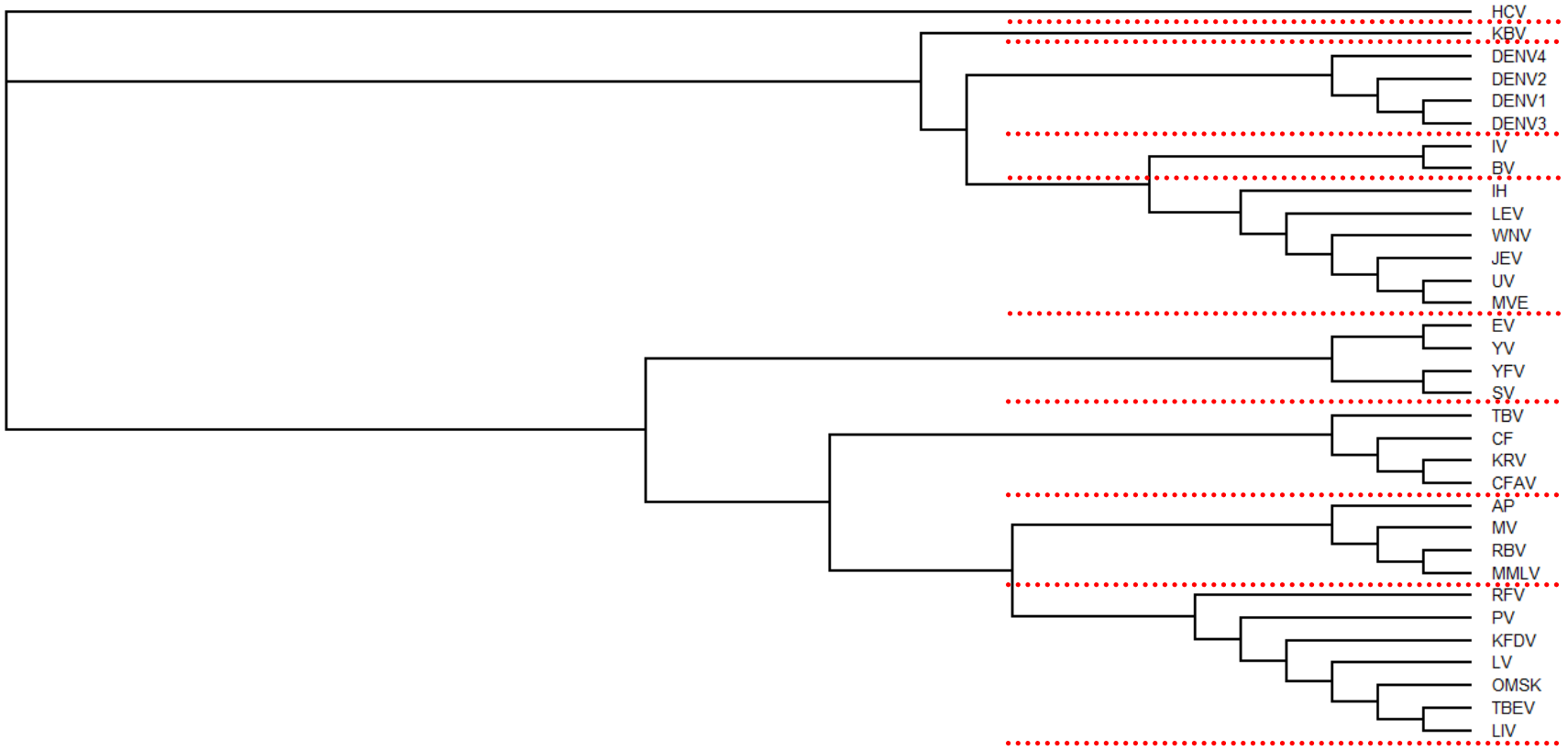
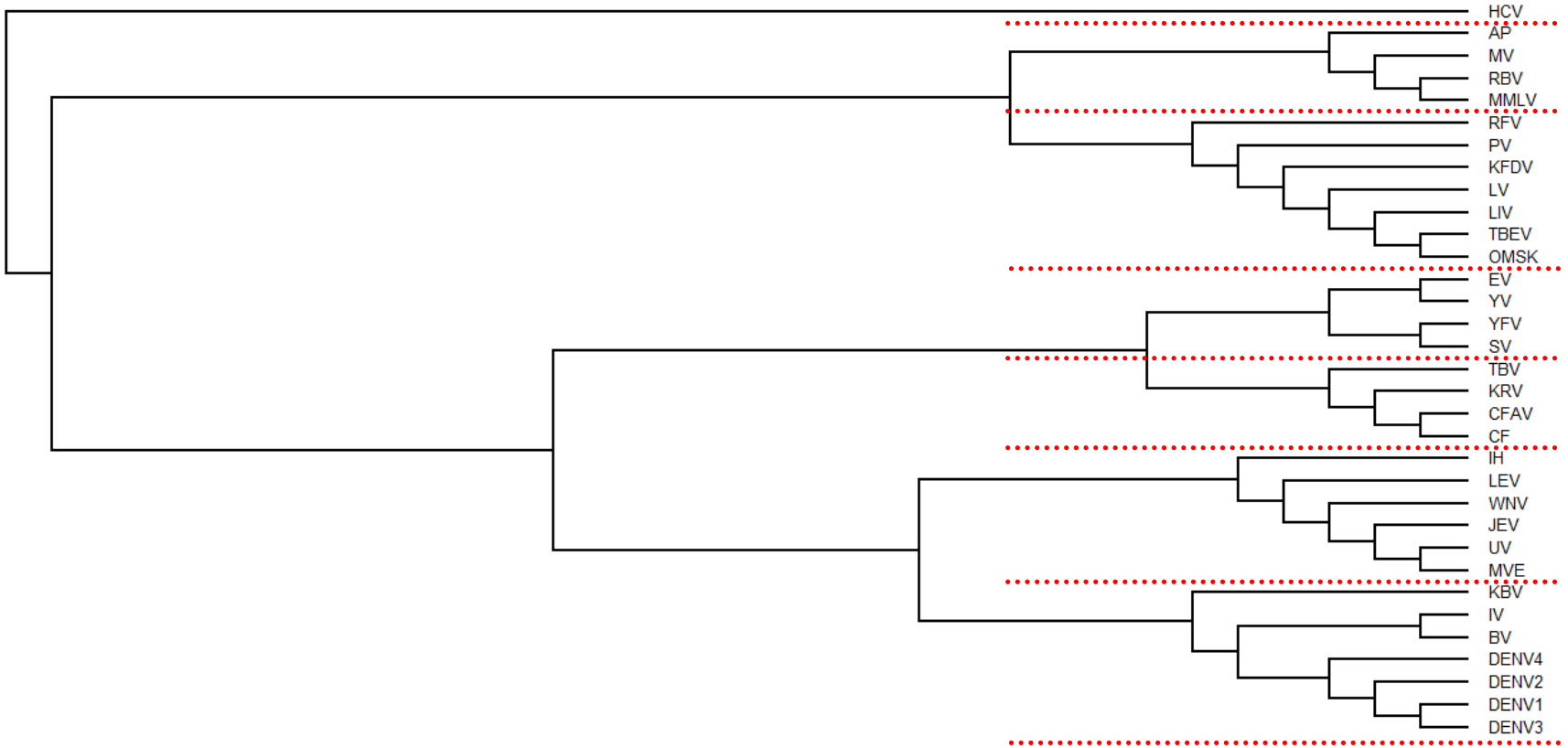


Figure 7.1: Phylogenetic relationship of polyprotein proteomes of selected 29 flaviviruses. The grouping of the viruses into phylogenetic clusters is indicated by the dotted red lines.

A) Envelope (E)


HCV
AP
MV
RBV
MMLV
RFV
PV
KFDV
LV
LIV
TBEV
OMSK
EV
YV
YFV
SV
TBV
KRV
CFAV
CF
IH
LEV
WNV
JEV
UV
MVE
KBV
IV
BV
DENV4
DENV2
DENV1
DENV3

B) NS3

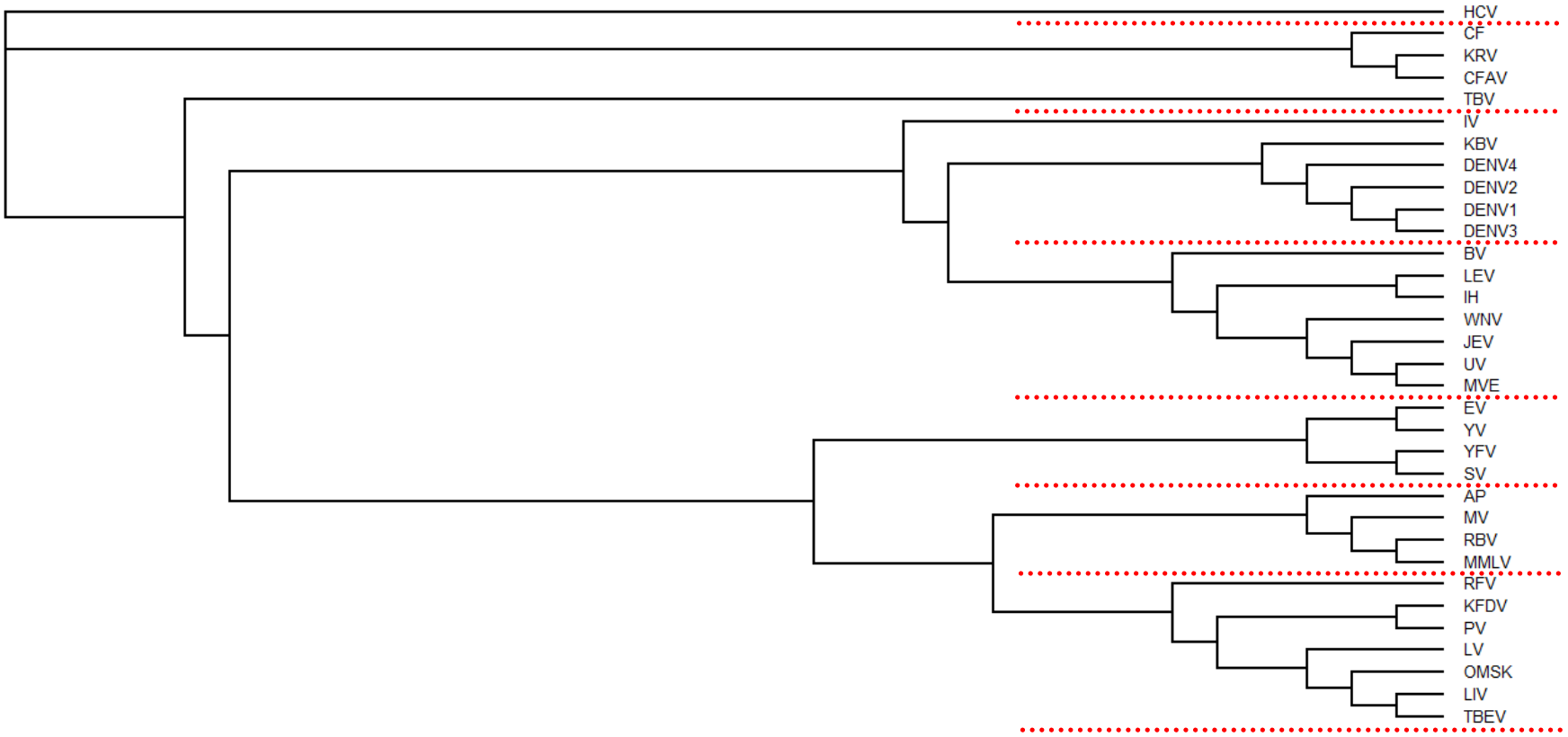
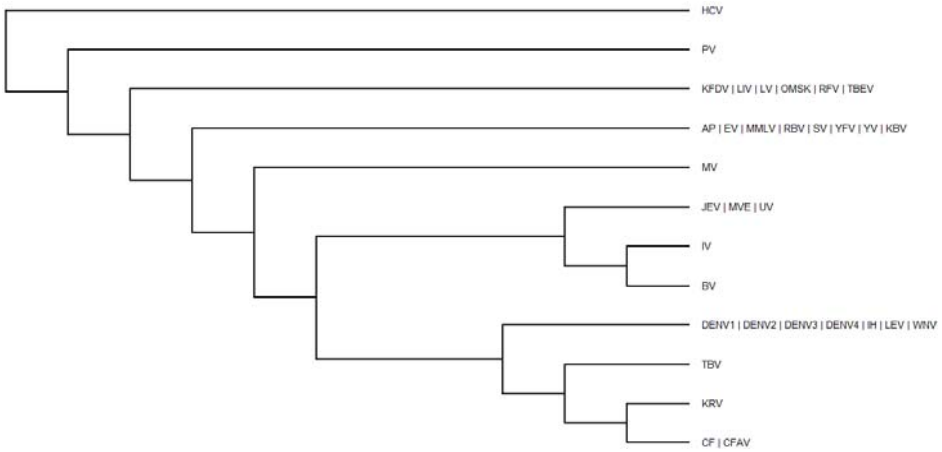
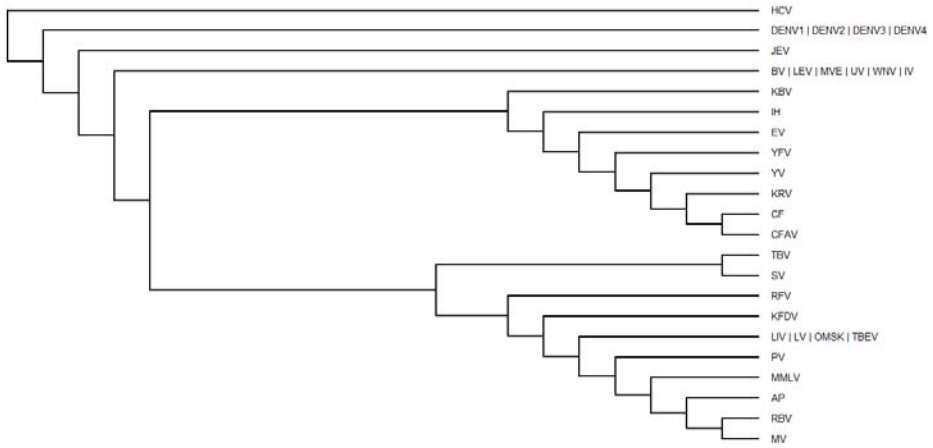


Figure 7.2: Phylogenetic relationship of A) the highly diverse envelope and B) the highly conserved NS3 protein of selected 29 flaviviruses. The grouping of the viruses into phylogenetic clusters is indicated by the dotted red lines. See Appendix 9 for the results of the other proteins analysed.

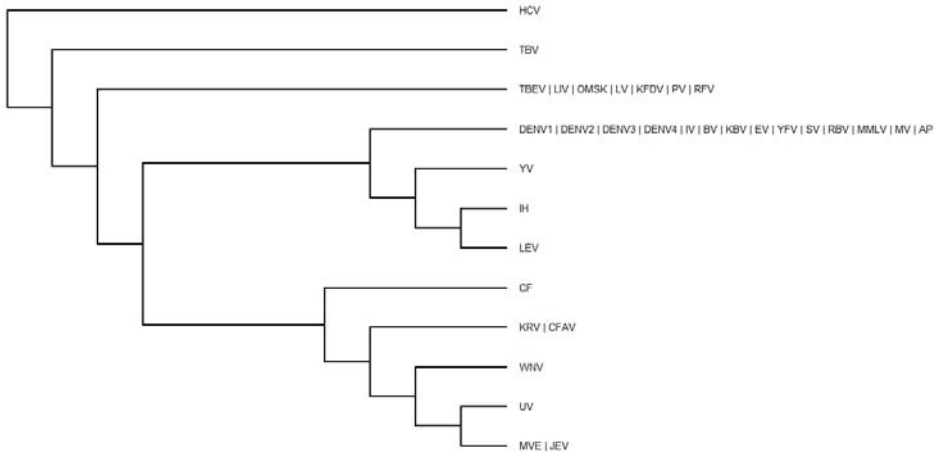
E $_{97}$VDRGWGNGCGLFGKG$_{111}$



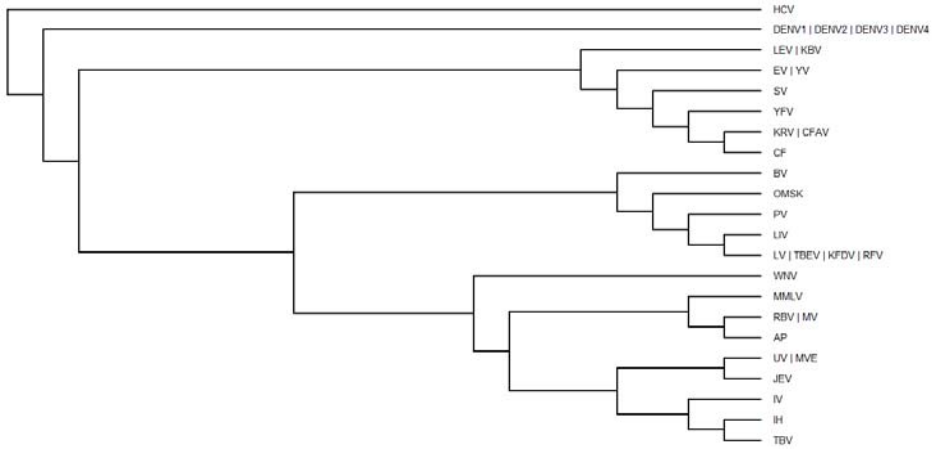
E $_{252}$VLGSQEGAMH$_{261}$



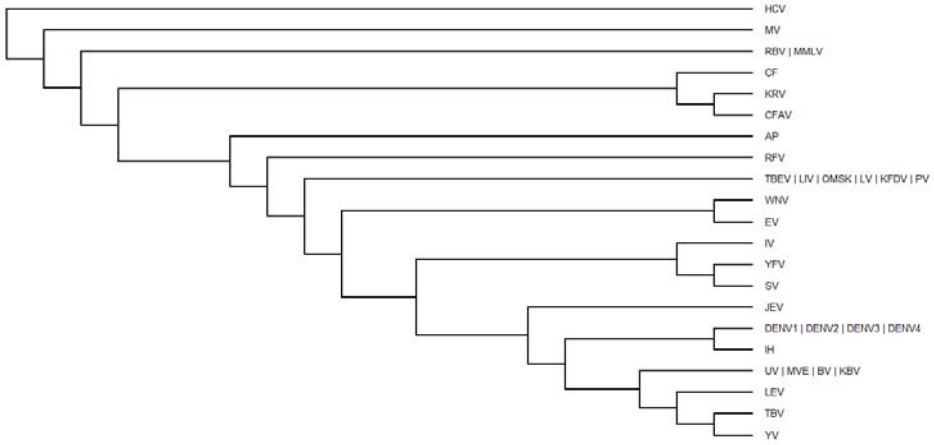
NS3 $_{46}$FHTMWHVTRG$_{55}$



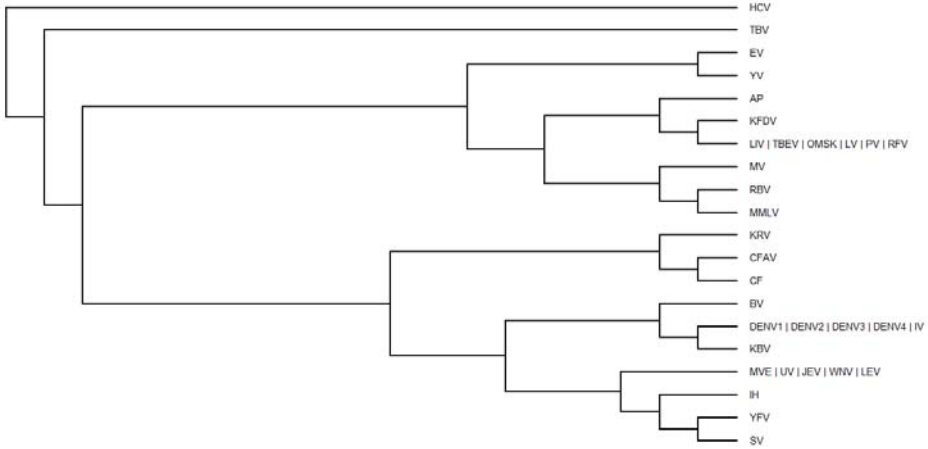
NS3 $_{148}$GLYGNGVVT$_{156}$



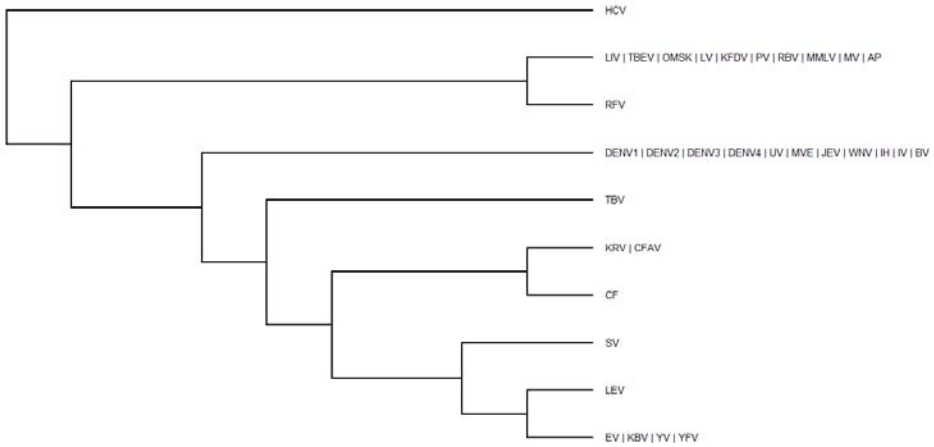
NS3 $_{189}$LTIMDLHPG$_{197}$



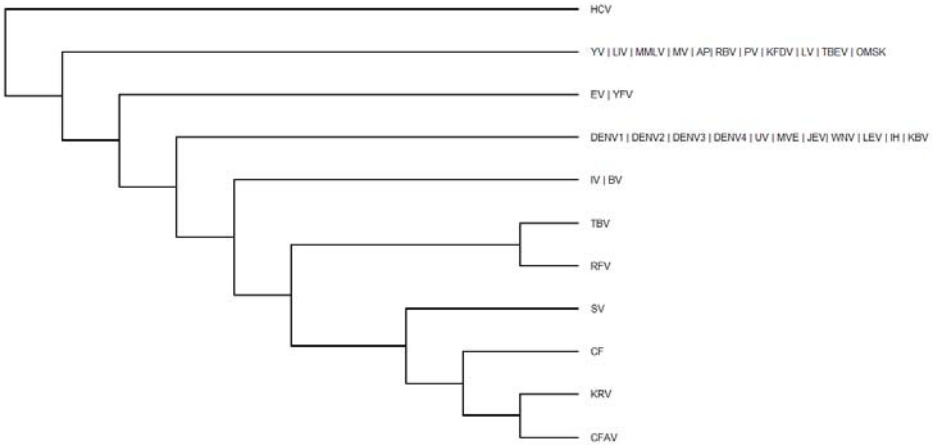
NS3 $_{256}$EIVDLMCHATFT$_{267}$



NS3 $_{284}$MDEAHFTD$_{P292}$



NS3 $_{313}$IFMTATPPG$_{321}$



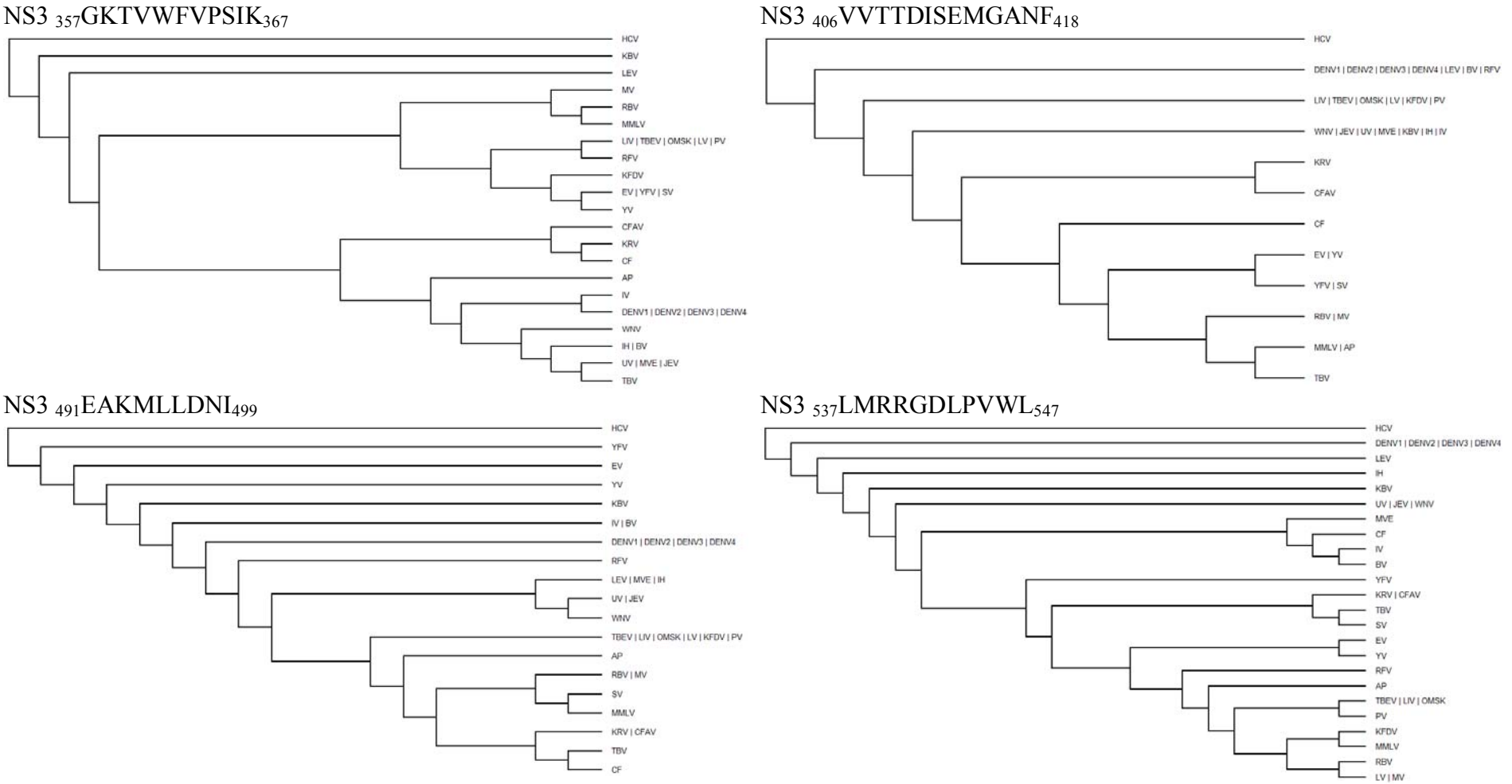
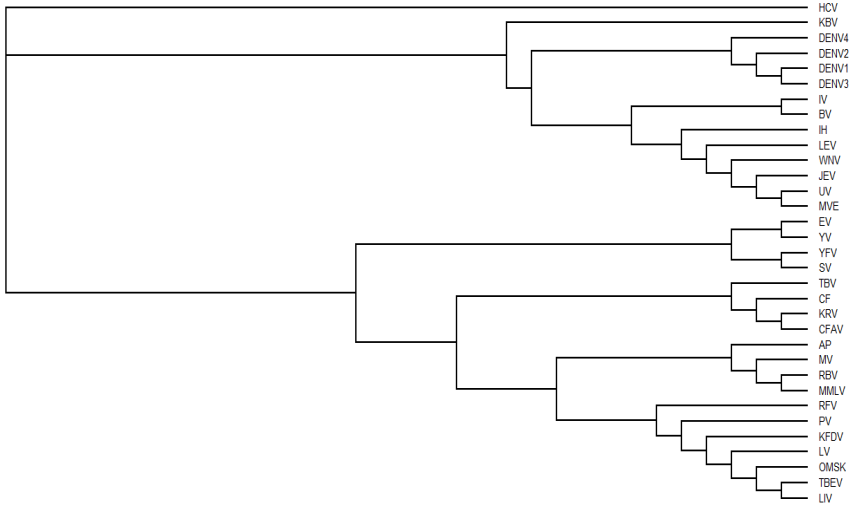
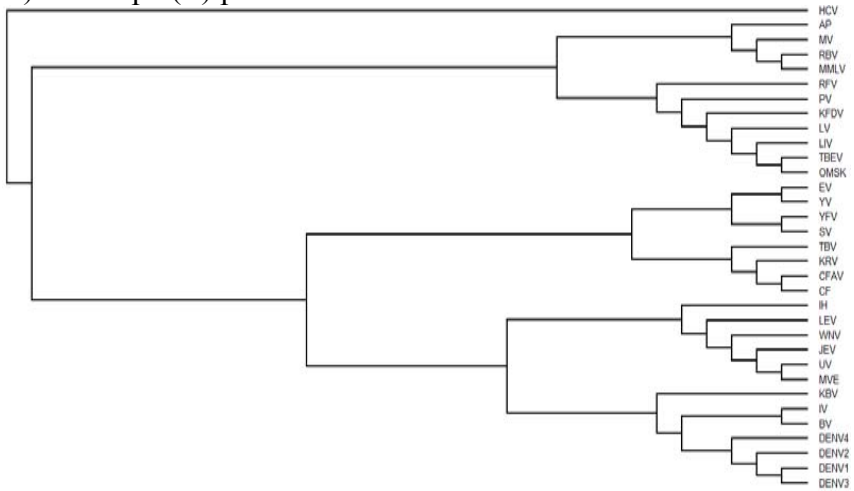


Figure 7.3: Phylogenetic relationship of *PE*s across selected flaviviruses. Only *PE*s from the envelope (E) and NS3 proteins are shown. See Appendix 9 for the results of the *PE*s from other proteins.

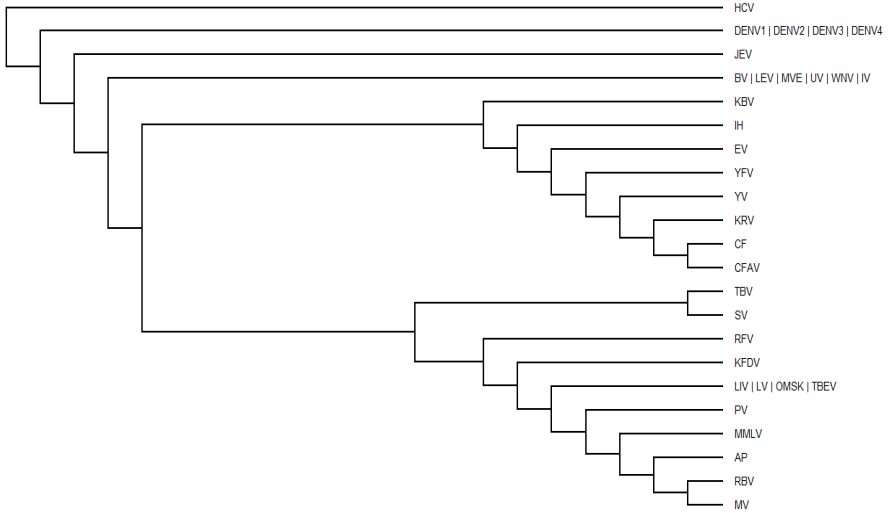
A) Proteome



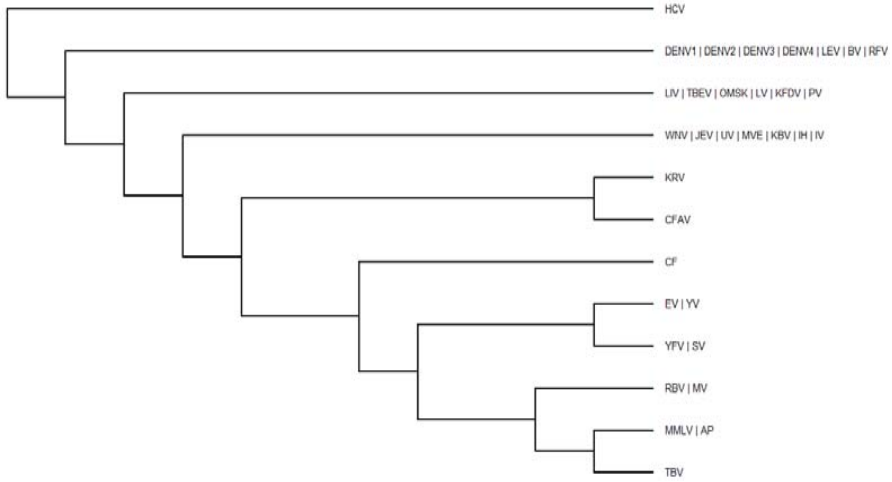
B) Envelope (E) protein



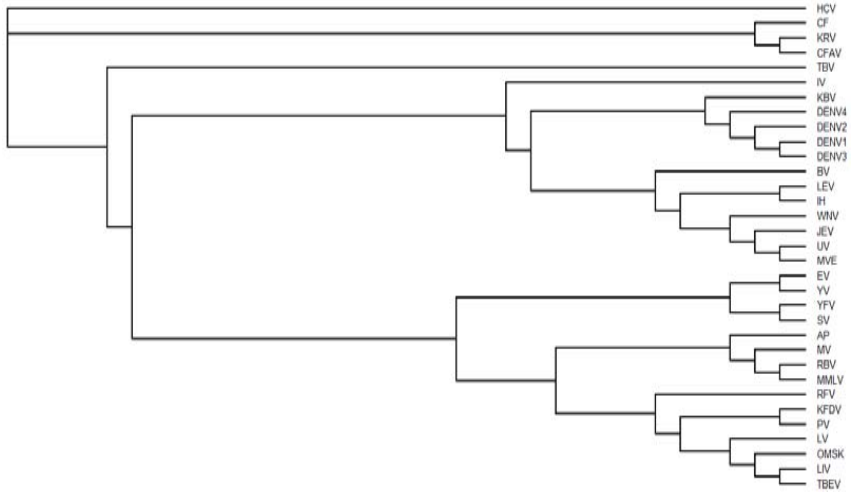
D) *PE* (E $_{252}$VLGSQEGAMH$_{261}$)



F) *PE* (NS3 $_{406}$VVTTDISEMGANF$_{418}$)



C) NS3 protein



E) *PE* (NS3 $_{537}$LMRRGDLPVWL$_{547}$)

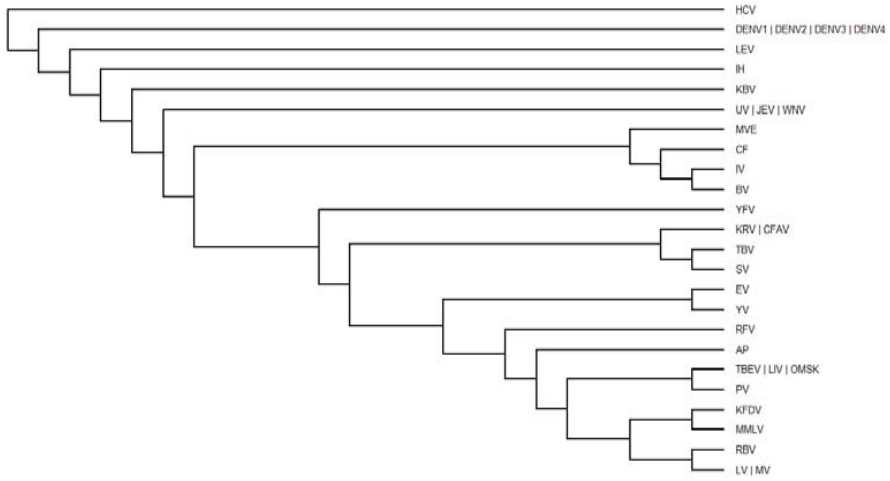


Figure 7.4: Differences in evolutionary relationships across the proteome, protein, and *PE* groupings.

7.4 Discussion

The phylogenetic trees at the proteome and protein levels indicated that the evolutionary patterns across flaviviruses are generally consistent. However, the analyses at the *PE* levels revealed that this is not true for the *PE* sites. The phylogenetic analysis of *PE*s across different flaviviruses revealed that different evolutionary pressures are likely to be acting upon the viruses (as shown by clustering of flaviviruses into multiple phylogenetic groups for the same *PE* site) and this pressure appears to be different for *PE*s from different regions of the protein (as illustrated by variation of the clustering patterns between *PE* sites). This suggests that for flaviviruses, the pattern of evolution of *PE*s between the viruses is generally different, despite sharing a common ancestral origin, genomic architecture and functional/structural role. This is probably in response to the adaptation or fitness of each virus to the different vector-host interaction environment.

The results emphasize the great complexity of conservation patterns of *PE*s in flaviviruses. Generally proteomes and proteins share similar overall properties. However, specific patterns of peptide conservation are not shared to the same extent and the patterns of similarity vary from peptide to peptide. An implication of this to peptide-based vaccine design is that pan-*Flavivirus* vaccine is unlikely. This is mainly because for a given *PE* sequence, viruses that are not part of the zero antigenic distance group (i.e. do not share the exact same *PE* sequence) but are closely related by a short antigenic distance, represent potential contributors of altered peptide ligands, upon co-infection, or secondary infection following vaccination or primary infection of the virus of interest. Further, *PE*s that exhibit zero antigenic distance across majority of the flaviviruses are rare, and those that do show extensive virus coverage (breadth) generally do not have the required depth (high representation of

the *PE* in each of the viruses shared). This reduces the number of *PE*s with good breadth and depth for experimental validation; this can be too small a starting number to guarantee success in the subsequent selections steps of vaccine design. It can be argued that a combination of *PE*s with both high breadth and depth can be utilized to cover the antigenic diversity of all the flaviviruses, however, because the individual *PE*s will still have viruses with short antigenic distance, the potential for altered ligand effect only multiplies. This could be further compounded by the inconsistent evolutionary pattern between *PE*s, suggesting unpredictable dynamics of evolution of *PE* sequences in future strains of the flaviviruses. All these suggest development of vaccines specific for each *Flavivirus* species and point to the direction of exploring species-specific *PE* sequences as peptide-based vaccine targets. Our pipeline (Chapter 5) is ideal for the identification and characterization of such targets.

7.5 Chapter summary

***Background*:** The family Flaviviridae consists of evolutionary related flaviviruses, including DENV and WNV, with common ancestral origin. Our antigenic diversity analysis of both DENV and WNV revealed extensive conservation of the majority of the *PE*s of each virus across numerous members of the genus *Flavivirus*. However, because this analysis of *PE*s only considered the flaviviruses that shared the specific *PE*, it does not provide a holistic view of the conservation pattern of a given *PE* to evolutionary related flaviviruses that share or do not share the *PE*. Complementing the analysis of depth and breadth, the evolutionary relationship of *PE*s across flaviviruses can be studied to better understand their conservation pattern and assess the possibility of a pan-*Flavivirus* vaccine. In this study, the author utilized evolutionary distance to investigate conservation pattern of *PE*s across flaviviruses. In

addition, this was benchmarked against at the proteome and protein level to assess for the stability of the relationship observed at the *PE* level.

***Results*:** Evolutionary analysis of DENV and 28 other flaviviruses was performed at the proteome, protein and *PE* levels using phylogenetic approach. Flaviviruses are ideal for this comparative analysis because they have similar genomic architecture and code for the same 10 proteins. Based on the phylogenetic tree of the proteome sequences of the selected 29 flaviviruses (including DENV), they grouped into eight clusters. However, analysis at the protein level revealed that there were differences in the clustering pattern. Nevertheless, the clustering pattern at the proteome and the protein level were generally consistent. At the *PE* level, the clustering patterns of the flaviviruses were significantly different from those observed at the proteome and protein levels. The flaviviruses clustered into multiple phylogenetic groups for the same *PE* region and the patterns varied between the *PE* sites. This suggests that for flaviviruses, the pattern of evolution of *PE*s between the viruses is generally different, despite sharing a common ancestral origin, genomic architecture and functional/structural role. This is probably in response to the adaptation or fitness of each virus to the different vector-host interaction environment.

***Conclusions*:** The results emphasize the great complexity of conservation patterns of *PE*s in flaviviruses. Generally proteomes and proteins share similar overall properties. However, specific patterns of peptide conservation are not shared to the same extent and the patterns of similarity vary from peptide to peptide. An implication of this to peptide-based vaccine design is that pan-*Flavivirus* vaccine is unlikely and suggests development of vaccines specific for each *Flavivirus* species, preferentially selecting the species-specific *PE* sequences. Our pipeline is ideal for the identification and characterization of such targets.

Chapter 8 General Discussions, Conclusions and Future Work

8.1 Antigenic diversity and implications for vaccine design

Reverse vaccinology, a bottom-up genomic approach, has been successfully applied to the development of vaccines against pathogens that were previously not suited to such development (Vernikos, 2008; Rappuoli and Covacci, 2003; Rappuoli, 2000). The pre-requisite for this approach is the sequence data of the target pathogen, which acts as input to various bioinformatics algorithms for prediction of putative antigens that are likely to be successful vaccine targets. These candidates can then be validated by a small number of key experiments in the lab. The approach has been successfully applied to the development of universal vaccines against group B *Streptococcus* (Maione *et al.*, 2005) and vaccine candidates against MenB (Pizza *et al.*, 2000), among others (Rappuoli and Covacci, 2003). Reverse vaccinology is a promising method for the high-throughput discovery of candidate vaccine targets that have the potential to mirror the dynamics and antigenic diversity of the target pathogen population, which includes the diversity of the interacting partner, the immune system. However, a big challenge to this end is the need to understand how vaccine developers can cover antigenic diversity, and develop a systematic approach to rationally screen pathogen data to select candidate vaccine targets that cover the diversity.

This thesis provides important insights into methods for covering antigenic diversity and is a significant contribution to the field of reverse vaccinology as it provides a pipeline to systematically screen and analyse pathogen data for peptides that cover antigenic diversity (*PE*s) prior to experimental validation. The pipeline helps efficiently deal with the astronomical combinatorial diversity possible between the numerous pathogen sequences and the highly polymorphic HLA binding partners of the immune system.

In this thesis, to better our understanding of antigenic diversity and explore ways to cover this diversity through antigenic peptides, the author utilized a systematic bioinformatics approach to study the relationship between genetic and antigenic diversity and examined the effect of sequence determinants, such as length and number of antigens, to antigenic diversity. Based on the insights gained, the author defined the criteria for peptides that cover antigenic diversity (*PE*). A large-scale systematic analysis was then performed to identify and characterize *PE*s from the large DENV sequences data available in public databases (>12,000). The methodology employed for the identification and characterization of *PE*s in DENV was then formulated as a generic systematic bioinformatics pipeline for similar analysis of other viruses. The pipeline was applied to WNV (and a number of other viruses) to demonstrate the generic nature and usefulness of the pipeline to flaviviruses and the results of WNV were used to perform a comparative analysis of its *PE*s to DENV. Further, conservation pattern of sequences across 28 flaviviruses corresponding to the DENV *PE*s were analysed. This was done to assess the stability of the conservation of the *PE*s across the flaviviruses, which share a common ancestral origin, genomic architecture and functional/structural role (Solomon and Mallewa, 2001; Henchal and Putnak, 1990).

Comprehensive DENV sequence data was collected, filtered of errors, and grouped before any analysis was performed. A computational method was developed for analysis of antigenic diversity of T-cell epitopes in DENV. The several advantages of the method include i) simple metric utilized ii) applicable for analysis of large number of sequences, either complete or partial and iii) most importantly, it does not require multiple sequence alignment of the sequence data, which can be difficult to achieve without misalignments for highly variable sequences. Through the application

of the method to DENV sequence data, the author demonstrated that complete coverage of antigenic diversity can be achieved by focusing on short regions of proteins. Covering the complete antigenic diversity in the context of all the possible different HLA molecules in the human population requires a large number of full-length protein sequences, which is not practical for the purpose of vaccine formulation using the traditional whole immunogen or the subunit vaccine strategy. Therefore, for us to focus on short conserved peptides and still cover the HLA polymorphism, we have to address the issue in a divide and conquer approach in the context of peptide-based vaccine strategy by identifying conserved peptides that are promiscuous to multiple HLA alleles (HLA supertype restricted epitopes). According to Sette *et al*., (1999), such targets for six or more of the major HLA supertypes are sufficient to capture the HLA polymorphism of nearly the whole human population.

To the best knowledge of the author, this thesis provides the first large-scale and complete identification and characterization of *PE*s of the four DENV serotype sequences (the pan-DENV sequences). Experimental validation in HLA transgenic mice and correspondence to reported T-cell epitopes immunogenic in humans provide evidence that the *PE*s are immunologically relevant and gives an indication of the reliability of the predictions for presence of promiscuous T-cell epitopes in the *PE*s.

The author proposes that *PE*s present in viral genomes are immutable functional tags, unlikely to change in the future, as indicated by the i) low peptide entropy they exhibited ii) their relevance to critical structure and function iii) extensive conservation with other genus members, such as other flaviviruses. This evolutionary stability of *PE*s over their entire known history makes them highly attractive candidates for therapeutic, prophylactic and diagnostic purposes, potentially

effective against a broad spectrum of the pathogen variants, including both existing and yet to emerge.

One of the features of DENV is the extensive protein sequence variability between the serotypes, and differences between similar amino acid sequences of T-cell epitopes recognized by the same HLA molecule in the event of multiple *Flavivirus* infection, or between a vaccine and subsequent pathogen challenge, are hypothesized to have an important role in the development of pathological immunity (Rothman, 2004). In this context, the *PE*s exhibiting none to almost negligible variant representation (0 to 5%) are likely to subvert pathological immunity. This negligible variant representation means that the probability of the immune system of an individual meeting a variant sequence of the four serotypes is very low, and, thus, minimizing the probability of altered peptide ligand effect potentially resulting in deleterious immune response. This feature of *PE*s provides another reason for their consideration in vaccine design. However, potential variants could also originate from flaviviruses that share the *PE*s, following co-infection or vaccination and secondary infection by these viruses homologous to DENV. It is, therefore, essential to assess the risks of occurrence of very similar sequences in other pathogens for the purpose of vaccine formulation. The comparative analysis of the distribution of *PE*s in nature serves as a starting point in this direction. Selecting *PE*s that are pathogen specific is one way of reducing the altered peptide ligand effect potential from other viruses.

The global bioinformatics and experimental approaches described in this thesis for DENV proved generic and useful to other flaviviruses as it was successfully applied. Thus, the generic approach can be used as a pipeline for similar large-scale analysis of other viruses. This is a significant contribution to the field of reverse vaccinology as it enables further filtering and analyses of ORFs identified *in silico*

from pathogen genomic data to select for *PE*s prior to experimental validation, thus, greatly reducing the efforts and cost of experimentation, while providing for systematic screening.

Application of the pipeline to other viruses enables comparative analysis of the characteristics of *PE*s between viruses. For example, it allowed comparison of the *PE*s of different viruses in the multi-dimensional context of i) number of *PE*s, ii) future conservation potential (entropy analysis), iii) conservation breadth - number of other viruses sharing the exact sequence of the *PE* (*i.e.* at least nine consecutive amino acids threshold used herein), iv) conservation depth - frequency or representation of the shared *PE* sequence in all known sequences of the corresponding virus, v) functional-structural relevance, vi) altered ligand potential by variants of the virus of interest or other viruses that share the *PE*, and vii) immunogenicity potential. Such comparative analysis will contribute to better understanding of *PE*s across pathogens and may provide insights into better design of vaccine strategies.

The conservation pattern analysis of *PE*s across different flaviviruses showed that the pattern of evolution of *PE*s between the viruses is complex, despite them sharing a common ancestral origin, genomic architecture and functional/structural role. This is probably in response to the adaptation of the virus to the different vector-host interaction environment. Implications of this to T-cell epitope peptide-based vaccine design include i) that pan-*Flavivirus* vaccine is not likely feasible, and ii) that for a particular *PE*, members of the same cluster that are not identical to each other but closely related (not sharing zero antigenic distance), represent potential contributors of altered peptide ligands. Our results indicate that vaccines need to be developed specific for each *Flavivirus* species, and that species-specific *PEs* are

attractive targets for research in this direction, which can be identified by application of our pipeline.

8.2 Strategies for dengue vaccine development

Dengue infection is a major worldwide medical problem, potentially affecting more than three billion people in more than 100 countries. No vaccine is currently licensed for human use. A vaccine is, therefore, urgently needed to lessen the global dengue disease burden. A successful dengue vaccine must be capable of simultaneously inducing a high level of long-lasting immunity to all four serotypes, to reduce the risk of potentially fatal DHF/DSS. Current dengue vaccine research focuses on several strategies (Hatch *et al.*, 2008; Whitehead *et al.*, 2007), such as live-attenuated or inactivated viruses, infectious clone-derived vaccines, immunogens vectored by various recombinant systems, subunit immunogens and DNA vaccine. The major focus is on the use of live-attenuated and infectious clone-derived vaccines. The other strategies, such as DNA and subunit vaccines, that focus on key parts of the pathogen are in early stages of development. Overall, the results to date appear inconclusive yet encouraging (Hatch *et al.*, 2008; Wilder-Smith and Deen, 2008; Whitehead *et al.*, 2007).

Developing a vaccine against DENV is a challenging task because of the antigenic diversity of the virus and the diversity of the immune system within the human host population. Understanding variation, both in the immune system and DENV, is necessary for dengue vaccine formulation. Current dengue vaccine strategies aim to cover the diversity between the four serotypes through simultaneous immunization with single strains of the four serotypes. However, utilizing a single strain of the four serotypes does not guarantee coverage of the diversity within and

between the serotypes, in particular when it is not known which of the epitopes in each of the strains will be targeted and recognized by the immune system. Recent studies even caution against the dogma that immunity to one homologous DENV serotype is protective against all subtype variants of the serotype (Zulueta *et al.*, 2006; Blaney *et al.*, 2005; Endy *et al.*, 2004). Studies with antibodies against DENV-3 showed that they do not always neutralize all known subtype variants of the serotype; similar results are expected for T-cell epitopes. Currently, methods for rational selection of dengue strains and antigens, which is crucial for successful vaccination (Boggiano *et al.*, 2005; Duffy *et al.*, 2005; Gaschen *et al.*, 2002), are not well established (Innis and Eckels, 2003). Further, the candidate vaccine antigens are not selected on the basis of possessing targets of immune responses that are recognized in the context of HLA supertypes. Thus, the large diversity of the human immune system at the population level may limit the effectiveness of the vaccines developed to certain proportion of the population only.

To tackle these issues, based on the results observed in this thesis, the author proposes peptide- or subunit-based vaccine approach as an ideal strategy to cover the diversity of both the virus and the human immune system. This involves focusing on short segments of the virus that are conserved, virus-specific and contain promiscuous T-cell epitopes (in the context of HLA supertypes), in addition to utilizing conserved neutralizing antibody epitopes, instead of using the "natural" form of the pathogen or selecting protein components of the pathogen. DNA vaccines using multi-epitope approach (Sette and Fikes, 2003) produced by recombinant technology are suitable for this purpose.

A commonly cited concern associated with the use of conserved epitopes is that they are generally reported to be poorly immunogenic (Wilson *et al.*, 2003; Parra

et al., 2000) and non immunodominant (Delves *et al.*, 1997). Immunogenicity can be improved by using adjuvants (Petrovsky and Aguilar, 2004) and high-affinity HLA-binding peptides, which have proven to be highly predictive of immunogenicity in the case of several viral pathogens (Altfeld *et al.*, 2001; Alexander *et al.*, 1997; Doolan *et al.*, 1997; Sette *et al.*, 1994). To address the immunodominance issue, it is important that epitope selection is not restricted to completely conserved sequences; instead, selection should include highly conserved sequences, such as those that exhibiting ≥ 80% representation in the known data, in order to include epitopes that are slightly variable, since immunodominant epitopes are generally localized in the variable regions of a virus. In addition, a selection bias toward highly conserved epitopes for peptide-based vaccines might offset the lack of immunodominance of conserved epitopes because it is thought that the absence of variable immunodominant epitopes may enhance the immunodominance of those that are conserved (Delves *et al.*, 1997). Antigenic drift due to strong immunologic pressure is also a concern with the peptide-based vaccine approach (Thomas *et al.*, 2006). A single amino acid change, even involving a conservative substitutions, can abolish recognition by T-cells (Sloan-Lancaster and Allen, 1996). To mitigate this effect, immunization strategies should be designed to induce cellular immune responses against multiple conserved epitopes of the virus.

Vaccine development is a complex process that includes data analysis, design of vaccine components, safety and efficacy determination and clinical trials. However, understanding the available data is an important first step that has proved to be increasingly critical to the success of subsequent steps. This work on dengue informatics represents an effort to complete the first step of virus sequence analysis in a systematic manner.

8.3 Vaccine informatics and future vaccines

Vaccines have proved to be among most powerful medical interventions. However, traditional vaccination strategies have clearly showed their limits for the development of effective vaccines against a number of disease agents. Recent advances in immunology and in genomics of disease agents now allow for rational design of genetically defined vaccines. These advances have led to the generation of large amount of data related to the immune system and of disease-agents. Nevertheless, currently available data represents only a tiny fraction of the natural library and data continues to accumulate at an exponential rate (Brusic and Petrovsky, 2003). The exponential growth of vaccine related data has created a need for better data management and analysis. Vaccine informatics provides a means for systematic study of large number of vaccine related data, enables selection of key experiments and facilitates experimental design. It is a practical science applied to the quest for designing new vaccines with the focus on bioinformatics-driven acquisition, manipulation and analysis of data related to the immune system and causative agents of diseases, such as viruses, bacteria, parasites and tumor, among others (Figure 8.1).

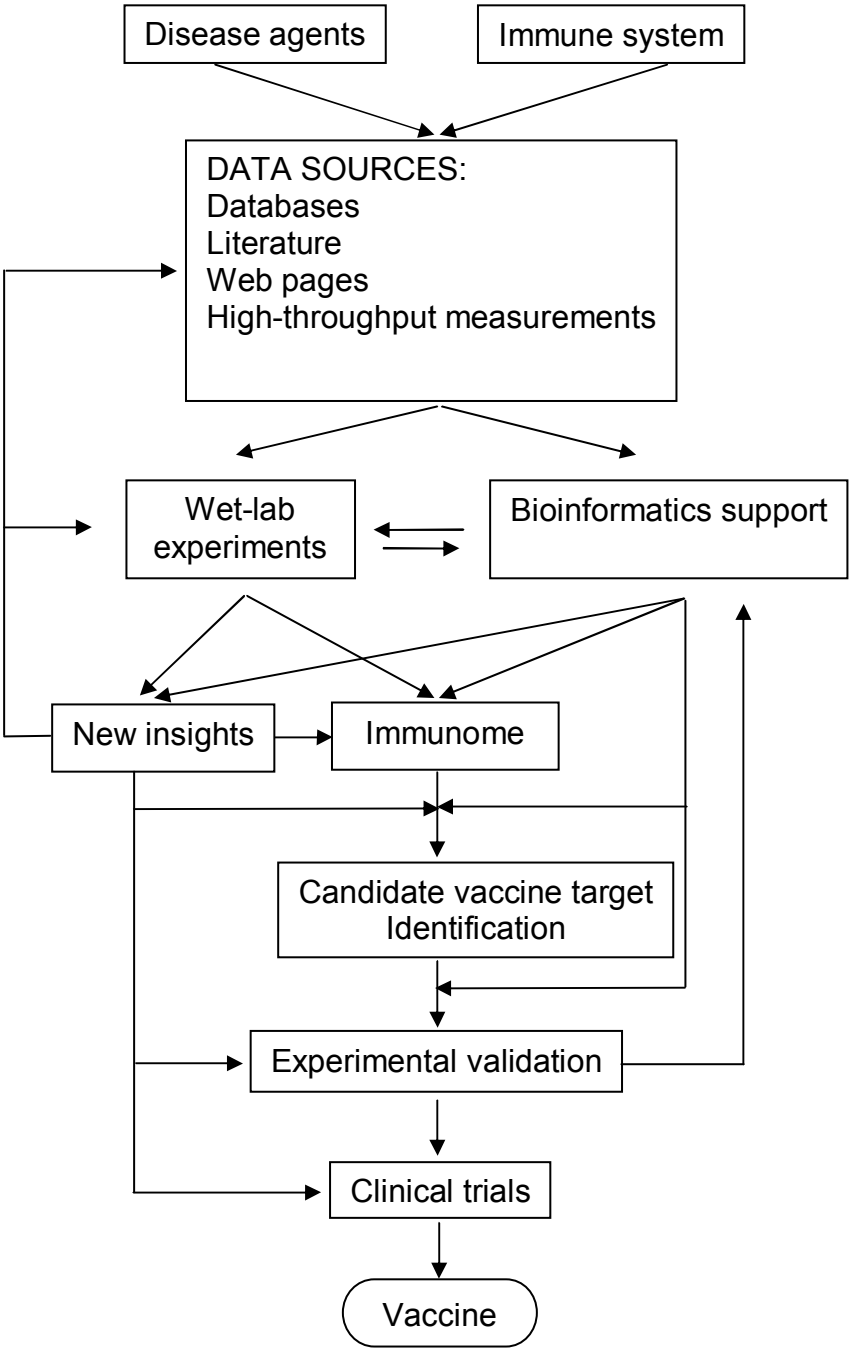


Figure 8.1: Vaccine informatics research. Large amount of data from genomics, proteomics and functional studies of the immune system and disease agents are stored in various data sources. Bioinformatics approaches are used to support collection, management, and systematic analysis of the data (including simulation of processes, generation of hypothesis and experimental design to test the hypothesis). The use of bioinformatics approaches in combination with experimental validation accelerates the discovery and facilitates better understanding of the components of the immunome. This may lead to new insights that may be utilized to enrich the data sources, better understand the immunome, aid in identification of vaccine targets, or clinical applications. Computational supports are used to pre-screen vaccine targets from the immunome and facilitate subsequent experimental design to test the targets. The results from the experiments performed may be feedback to the computational

models to further refine them. Candidate targets identified proceed into clinical trials and may eventually be used for vaccine production. Vaccine development, being a combinatorial problem, will benefit significantly from computational support.

Vaccine informatics, still in its infancy, already has the potential to revolutionize the process of vaccine design for the development of new, safe, and effective vaccines. The use of bioinformatics approaches in combination with experimental validation enable large number of laboratory experiments to be avoided, thus, accelerating vaccine research and diminishing discovery cost. Currently, major bioinformatics developments in vaccine research include support in managing and analysing large quantities of data, modeling of the immune system, analysing the diversity of disease-agents and complexity of the human immune system, and high-throughput screening of candidate vaccine targets. Future developments include advanced applications in addressing complex immunome-related problems, both at cellular and system level, and capitalizing on accumulated data, information and knowledge in vast repositories. Efforts are on the way to build virtual immune system by progressively adding together models representing each known facet of antigen presentation (Larsen *et al.*, 2005; Petrovsky *et al.*, 2003).

ImmunoGrid, an effort to simulate the processes of the mammalian immune system at the molecular, cellular and tissue levels using Grid technologies, is a step forward towards the development of a virtual immune system (www.immunogrid.org) (Pappalardo *et al.*, 2009). Doing so would be trying to match the complexity of the real immune system and this will greatly assist our understanding of not only normal functioning of the immune system but also will help elucidate how immunization affects immune function. Better understanding of the components of the human immunome offers a great promise for development of effective vaccines. Never

before the quest for deciphering the fine details of initiation, regulation and modulation of the immune responses has been based more on accumulated knowledge and less on hit-and-miss approach of yesteryears.

Future vaccines will be minimalistic in approach by focusing on key parts of the pathogen, such as regions containing epitopes that cover antigenic diversity and, thus, will target immunologically similar subgroups of the human population and multiple pathogen variants. This is evident from the trend observed in evolution of vaccine strategies, which has seen a shift from whole organisms to recombinant proteins, and further towards the ultimate in minimalist vaccinology, the peptide/epitope/multi-epitope. The minimalist approach is also expected to cover the safety concerns that are associated with the traditional vaccine approach of using whole organism (Sette and Fikes, 2003; Dertzbaugh, 1998). Vectored vaccines, suitable for 'combination immunization' that are produced by recombinant DNA technology and contain multivalent minimal antigens to protect against multiple infections, are considered to be the future of vaccinology (Kutzler and Weiner, 2008).

The author foresees that the future will bring increased integration of vaccine research with advances in immunology, molecular biology, genomics, proteomics, informatics, and high-throughput instrumentation, collective defined as the emerging field of "vaccinomics", which is hailed to be responsible for the next 'golden age' in vaccinology (Poland *et al.*, 2008). Awareness of the novel technological possibilities in vaccine research is also expected to grow. Future vaccinology will be based on detailed understanding of immune function, optimal stimulation of immune responses (using adjuvants) and precise mapping and rational selection of immune targets (Brusic *et al.*, 2005). To achieve this, vaccine development will routinely be conducted through large-scale functional studies supported by genomics, proteomics,

and informatics techniques prior to clinical trials. This will provide an increased range of immune targets for vaccine design. The author expects the emergence of new generation of vaccines to be personalised to both the genetic make-up of the human population and of the disease agents. In summary, vaccinology will experience rapid progress and will eventually deliver benefits to patients from improved diagnosis, treatment and prevention of diseases.

8.4 Conclusions

Bioinformatics is essential for the analysis and interpretation of complex and large quantity of biological data generated by functional studies and high throughput technologies. It is used to propose the next sets of experiments and, most importantly, to derive better understanding of biological processes. The number of viral sequence data in public databases is increasing rapidly. Experimental approaches to study this large data pool for the development of immune interventions, such as vaccines, against the viruses are time-consuming, costly and almost impractical. Through combination of bioinformatics and experimental approaches, it is possible to select key experiments and help optimize experimental design. Computer algorithms are increasingly used to speed-up the process of knowledge discovery by helping to identify critical experiments for testing hypothesis built upon the result of computational screening. A number of successful examples for application of computer models to study immunological problems have been described in (Brusic *et al.*, 2005). Such examples illustrate the power of computational approach to complex problems involving potentially vast datasets with potential biases, errors and discrepancies.

This thesis focused on a systematic bioinformatics approach to analyzing antigenic diversity of targets of cellular immune responses (T-cell epitopes) in reported sequences of the four DENV serotypes. Analysis of antigenic diversity presents us with a unique opportunity to improve our understanding of the immune response to viral variants and aid in identification of peptide targets for vaccine formulation. Comprehensive DENV sequence data was collected, filtered of errors, and grouped before any analysis. A simple, generic and systematic bioinformatics methodology developed was applied for the analysis of antigenic diversity of T-cell epitopes in DENV datasets for the proteins of the four serotypes. Antigenic diversity analysis showed that the number of unique protein sequences required to represent complete antigenic diversity of short peptides in DENV was significantly smaller than that required to represent complete protein sequence diversity. Short-peptide antigenic diversity showed an asymptotic relationship to the number of unique protein sequences, indicating that for large sequence sets (~200) the addition of new protein sequences has marginal effect to increasing antigenic diversity. A near-linear relationship was observed between the extent of antigenic diversity and the length of protein sequences, suggesting that, for the practical purpose of vaccine development, antigenic diversity of short peptides from DENV can be represented by short, conserved regions of sequences (~<100 aa) within viral antigens that are specific targets of immune responses (such as T-cell epitopes specific to particular HLA alleles), in particular promiscuous T-cell epitopes. This provided evidence that there are limited numbers of antigenic combinations in protein sequence variants of a viral species and that short, conserved regions of the viral protein are sufficient to cover antigenic diversity of T-cell epitopes. The methodology for analysis of antigenic

diversity has direct application to the analysis of other viruses, such as influenza A virus and human immunodeficiency virus (HIV).

Based on the insights gained from the analysis of antigenic diversity, the author identified and characterized DENV peptides that cover antigenic diversity (*PE*s) – conserved, short viral sequence fragments in the DENV proteome that are promiscuous T-cell epitopes. A large-scale identification and analysis of evolutionarily highly conserved amino acid sequences of the entire DENV proteome, with a focus on sequences of nine amino acids or more, and thus immune-relevant as potential T-cell epitopes was undertaken. Forty-four (44) pan-DENV sequences of at least nine amino acids were highly conserved and identical in 80% or more of all recorded DENV sequences, and the majority were found to be immune-relevant by their correspondence to known or putative HLA-restricted promiscuous T-cell epitopes. These sequences are potential *PE*s as they potentially cover both the diversity of the DENV and variations in immune system among individuals (HLA polymorphism). The conservation of these sequences through the entire recorded DENV genetic history suggests that they are likely to remain conserved in the future and supports their possible value for diagnosis, prophylactic and/or therapeutic applications.

The combination of bioinformatics and experimental approaches applied herein provides a novel pipeline for large-scale and systematic analysis of *PE*s of other pathogens, such as for rapidly mutating viruses, including influenza A virus and HIV. This approach provides an experimental basis for the design of pathogen specific, T-cell epitope peptide-based vaccines that are targeted to majority of the genetic variants of the pathogen, and are effective for a broad range of differences in HLAs among the global human population.

The generic nature and usefulness of the approach to flaviviruses was demonstrated through its customized application to WNV for identification and characterization of *PE*s, which allowed comparative analysis of the characteristics of *PE*s between pathogens of interest. In addition, conservation pattern analysis of the *PE*s of DENV with corresponding sequences of 28 other flaviviruses revealed complex pattern of evolution at the *PE* sites.

The work described in this thesis, application of bioinformatics to the first step in exploring the potential of sequence data for vaccine discovery, is a step forward for the field of reverse vaccinology as it enables the systematic screening and analyses of pathogen data in the context of the immune system, which would otherwise be impossible to carry out experimentally, due to the large combinatorial diversity possible between the numerous pathogen sequences and the highly polymorphic HLA binding partners. It therefore significantly reduces the efforts and cost of experimentation, while providing for systematic screening. We are entering a new era of vaccine immunomics synergistically powered by integration of informatics and other advances in immunology, molecular biology, genomics, proteomics, high-throughput instrumentation, providing for detailed understanding of immune function, optimal stimulation of immune responses (using adjuvants) and precise mapping and rational selection of immune targets that cover antigenic diversity. This all is expected to lead towards the development of new generation of vaccines, personalised to both the genetic make-up of the human population and of the pathogen. The author of this thesis believes that his contributions will convert into practical vaccine solutions and hopes that, soon, a novel vaccine formulation that can protect against DENV diseases will be available to the public.

8.5 Future work

In this thesis, to cover DENV antigenic diversity, peptide sequences common across the four serotypes and immunologically relevant as promiscuous epitopes (pan-DENV sequences) were identified and characterized. An advantage of these sequences is that they exhibit none to negligible variant representation. Further, utilizing pan-DENV sequences specific to the virus is beneficial than those that are not because of the low probability of contribution of altered peptide ligand effect from other flaviviruses.

An alternative strategy to the utility of pan-DENV sequences for covering of antigenic diversity is to identify conserved, serotype specific peptide sequences that contain HLA supertype-restricted epitopes. The rationale behind this strategy is that pan-DENV sequences still pose a risk for altered ligand effect, although minimal; only two pan-DENV sequences were completely conserved (100% representation) and had zero variants. T-cell epitope sequences that are specific to each serotype, are not likely to share cross-reactivity between them, and their high intra-serotype conservation will provide high coverage of its variants. Such epitope sequences are hypothesized to cover antigenic diversity of each serotype and yet avoid the possibility of potential variants resulting in altered ligand effect because of the significant sequence difference between the serotypes (Rothman, 2004). This approach has been suggested by few researchers (Mongkolsapaya *et al.*, 2006; Mangada and Rothman, 2005; Mongkolsapaya *et al.*, 2003; Kurane *et al.*, 1998) based on their experimental observation supporting the notion that highly or completely conserved, serotype-specific epitopes are attractive for dengue vaccine. This approach is particularly advantageous for highly diverse viruses, such as HIV, for which there are only few peptide sequences highly conserved across the different clade groups

(unpublished data from our group). The author plans to identify and characterize these sequences for DENV as part of his future research work.

The two contrasting strategies provide options to vaccine developers for the search of candidate vaccine targets. Given the large number of DENV sequence, the significant heterogeneity between the serotypes, and the diversity of the immune system, identifying such conserved, serotype-specific promiscuous epitopes experimentally is a challenging task. The author proposes application of the bioinformatics pipeline developed herein, which is suitable for such large-scale analysis, to systematically screen and select such peptide sequences for experimental validation. For the identification of conserved, serotype specific sequences, the AVANA component of the pipeline will be expanded to include the search methodology for such sequences (see example in Figure 8.2), in addition to the current pan-serotype approach.

Having identified all such peptides, the author then would analyse the frequency or representation of the sequences within the corresponding serotype dataset to filter out sequences not highly or completely represented within the serotype. The ensuing list of sequences would then be subjected to prediction of T-cell epitopes in the context of various HLA class I and II supertypes, by use of prediction algorithms described herein. The structural-functional relevance of the sequences and their conservation in other non-dengue viruses would also be examined using the methodology defined in this thesis. The final set of sequences will then be subject to experimental validation for their immunological relevance.

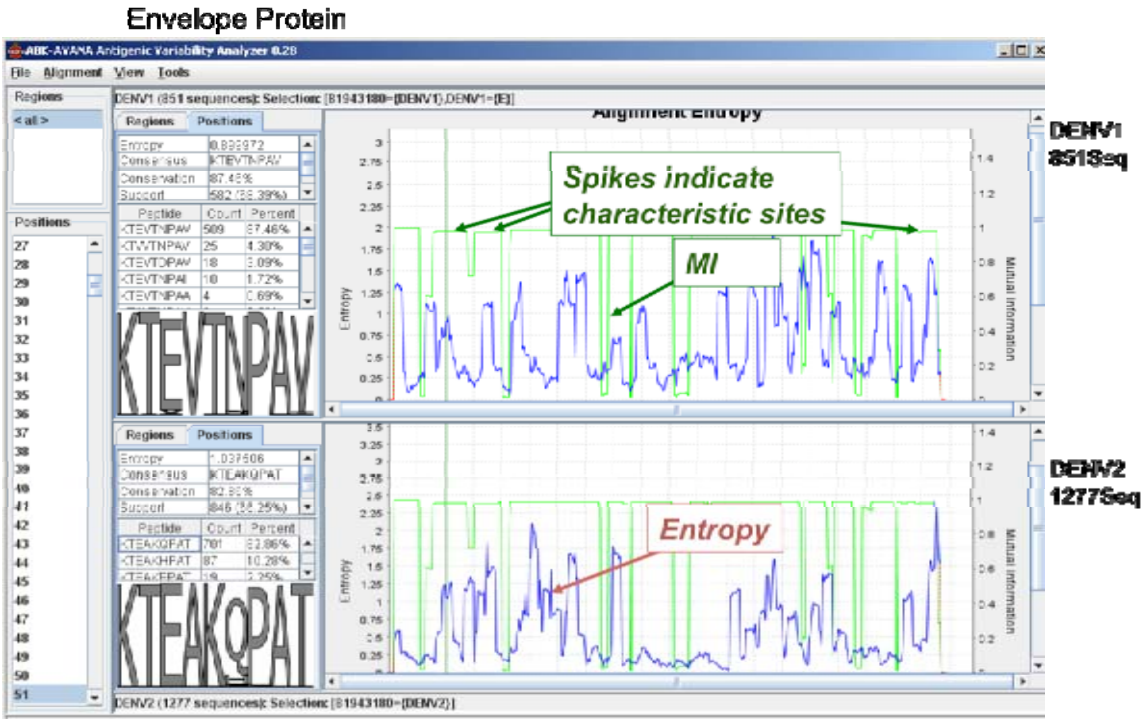


Figure 8.2: An example of application of AVANA to identify characteristic sites between sequence alignments of DENV-1 and DENV-2 envelope proteins. Spikes indicate sites with mutual information value of 1 and characteristic to the dataset. Mutual information is a measure of the relationship between two variables, and is derived by comparing the entropies of the datasets.

References

Acierno, P.M., Newton, D.A., Brown, E.A., Maes, L.A., Baatz, J.E. and Gattoni-Celli, S. (2003). Cross-reactivity between HLA-A2-restricted FLU-M1:58-66 and HIV p17 GAG:77-85 epitopes in HIV-infected and uninfected individuals, J Transl Med, *1*, 3.

Alexander, J., Oseroff, C., Sidney, J. and Sette, A. (2003). Derivation of HLA-B*0702 transgenic mice: functional CTL repertoire and recognition of human B*0702-restricted CTL epitopes, Hum Immunol, *64*, 211-23.

Alexander, J., Oseroff, C., Sidney, J., Wentworth, P., Keogh, E., Hermanson, G., Chisari, F.V., Kubo, R.T., Grey, H.M. and Sette, A. (1997). Derivation of HLA-A11/Kb transgenic mice: functional CTL repertoire and recognition of human A11-restricted CTL epitopes, J Immunol, *159*, 4753-61.

Allison, S.L., Schalich, J., Stiasny, K., Mandl, C.W. and Heinz, F.X. (2001). Mutational evidence for an internal fusion peptide in flavivirus envelope protein E, J Virol, *75*, 4268-75.

Altfeld, M.A., Livingston, B., Reshamwala, N., Nguyen, P.T., Addo, M.M., Shea, A., Newman, M., Fikes, J., Sidney, J., Wentworth, P., Chesnut, R., Eldridge, R.L., Rosenberg, E.S., Robbins, G.K., Brander, C., Sax, P.E., Boswell, S., Flynn, T., Buchbinder, S., Goulder, P.J., Walker, B.D., Sette, A. and Kalams, S.A. (2001). Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif, J Virol, *75*, 1301-11.

Bashyam, H.S., Green, S. and Rothman, A.L. (2006). Dengue virus-reactive CD8+ T cells display quantitative and qualitative differences in their response to variant epitopes of heterologous viral serotypes, J Immunol, *176*, 2817-24.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004). The Pfam protein families database, Nucleic Acids Res, *32*, D138-41.

Beaumier, C.M., Mathew, A., Bashyam, H.S. and Rothman, A.L. (2008). Cross-reactive memory CD8(+) T cells alter the immune response to heterologous secondary dengue virus infections in mice in a sequence-specific manner, J Infect Dis, *197*, 608-17.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2006). GenBank, Nucleic Acids Res, *34*, D16-20.

Berkhoff, E.G., Geelhoed-Mieras, M.M., Fouchier, R.A., Osterhaus, A.D. and Rimmelzwaan, G.F. (2007). Assessment of the extent of variation in influenza A virus cytotoxic T-lymphocyte epitopes by using virus-specific CD8+ T-cell clones, J Gen Virol, *88*, 530-5.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank, Nucleic Acids Res, *28*, 235-42.

Berzofsky, J.A., Pendleton, C.D., Clerici, M., Ahlers, J., Lucey, D.R., Putney, S.D. and Shearer, G.M. (1991). Construction of peptides encompassing multideterminant clusters of human immunodeficiency virus envelope to induce in vitro T cell responses in mice and humans of multiple MHC types, J Clin Invest, *88*, 876-84.

Bian, H. and Hammer, J. (2004). Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE, Methods, *34*, 468-75.

Billoir, F., de Chesse, R., Tolou, H., de Micco, P., Gould, E.A. and de Lamballerie, X. (2000). Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector, J Gen Virol, *81*, 781-90.

Biswas, T., Aihara, H., Radman-Livaja, M., Filman, D., Landy, A. and Ellenberger, T. (2005). A structural basis for allosteric control of DNA recombination by lambda integrase, Nature, *435*, 1059-66.

Bjorkman, P.J. and Parham, P. (1990). Structure, function, and diversity of class I major histocompatibility complex molecules, Annu Rev Biochem, *59*, 253-88.

Blaney, J.E., Jr., Matro, J.M., Murphy, B.R. and Whitehead, S.S. (2005). Recombinant, live-attenuated tetravalent dengue virus vaccine formulations induce a balanced, broad, and protective neutralizing antibody response against each of the four serotypes in rhesus monkeys, J Virol, *79*, 5516-28.

Boggiano, C., Moya, R., Pinilla, C., Bihl, F., Brander, C., Sidney, J., Sette, A. and Blondelle, S.E. (2005). Discovery and characterization of highly immunogenic and broadly recognized mimics of the HIV-1 CTL epitope Gag77-85, Eur J Immunol, *35*, 1428-37.

Bondre, V.P., Jadi, R.S., Mishra, A.C., Yergolkar, P.N. and Arankalle, V.A. (2007). West Nile virus isolates from India: evidence for a distinct genetic lineage, J Gen Virol, *88*, 875-84.

Brinton, M.A. and Dispoto, J.H. (1988). Sequence and secondary structure analysis of the 5'-terminal region of flavivirus genome RNA, Virology, *162*, 290-9.

Brinton, M.A., Kurane, I., Mathew, A., Zeng, L., Shi, P.Y., Rothman, A. and Ennis, F.A. (1998). Immune mediated and inherited defences against flaviviruses, Clin Diagn Virol, *10*, 129-39.

Brown, S.A., Stambas, J., Zhan, X., Slobod, K.S., Coleclough, C., Zirkel, A., Surman, S., White, S.W., Doherty, P.C. and Hurwitz, J.L. (2003). Clustering of Th cell

epitopes on exposed regions of HIV envelope despite defects in antibody activity, J Immunol, *171*, 4140-8.

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

Brusic, V., August, J.T. and Petrovsky, N. (2005). Information technologies for vaccine research, Expert Rev Vaccines, *4*, 407-17.

Brusic, V., Bajic, V.B. and Petrovsky, N. (2004). Computational methods for prediction of T-cell epitopes--a framework for modelling, testing, and applications, Methods, *34*, 436-43.

Brusic, V., Petrovsky, N., Zhang, G. and Bajic, V.B. (2002). Prediction of promiscuous peptides that bind HLA class I molecules, Immunol Cell Biol, *80*, 280-5.

Brusic, V. and Zeleznikow, J. (1999). Computational binding assays of antigenic peptides, Lett. Pept. Sci., *6*, 313-324.

Bui, H.H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K.A., Mothe, B.R., Chisari, F.V., Watkins, D.I. and Sette, A. (2005). Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications, Immunogenetics, *57*, 304-14.

Bukowski, J.F., Kurane, I., Lai, C.J., Bray, M., Falgout, B. and Ennis, F.A. (1989). Dengue virus-specific cross-reactive CD8+ human cytotoxic T lymphocytes, J Virol, *63*, 5086-91.

Chao, D.Y., King, C.C., Wang, W.K., Chen, W.J., Wu, H.L. and Chang, G.J. (2005). Strategically examining the full-genome of dengue virus type 3 in clinical isolates reveals its mutation spectra, Virol J, *2*, 72.

Clute, S.C., Watkin, L.B., Cornberg, M., Naumov, Y.N., Sullivan, J.L., Luzuriaga, K., Welsh, R.M. and Selin, L.K. (2005). Cross-reactive influenza virus-specific CD8+ T cells contribute to lymphoproliferation in Epstein-Barr virus-associated infectious mononucleosis, J Clin Invest, *115*, 3602-12.

Cope, A.P., Patel, S.D., Hall, F., Congia, M., Hubers, H.A., Verheijden, G.F., Boots, A.M., Menon, R., Trucco, M., Rijnders, A.W. and Sonderstrup, G. (1999). T cell responses to a human cartilage autoantigen in the context of rheumatoid arthritis-associated and nonassociated HLA-DR4 alleles, Arthritis Rheum, *42*, 1497-507.

Crochu, S., Cook, S., Attoui, H., Charrel, R.N., De Chesse, R., Belhouchet, M., Lemasson, J.J., de Micco, P. and de Lamballerie, X. (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes, J Gen Virol, *85*, 1971-80.

Cunha-Neto, E. (1999). MHC-restricted antigen presentation and recognition: constraints on gene, recombinant and peptide vaccines in humans, Braz J Med Biol Res, *32*, 199-205.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, Nucleic Acids Res, *34*, W362-5.

De Groot, A.S. (2004). Immunome-derived vaccines, Expert Opin Biol Ther, *4*, 767-72.

De Groot, A.S., Marcon, L., Bishop, E.A., Rivera, D., Kutzler, M., Weiner, D.B. and Martin, W. (2005). HIV vaccine development by computer assisted design: the GAIA vaccine, Vaccine, *23*, 2136-48.

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

De Groot, A.S. and Rothman, F.G. (1999). In silico predictions; in vivo veritas, Nat Biotechnol, *17*, 533-4.

De Groot, A.S., Saint-Aubin, C., Bosma, A., Sbai, H., Rayner, J. and Martin, W. (2001). Rapid determination of HLA B*07 ligands from the West Nile virus NY99 genome, Emerg Infect Dis, *7*, 706-13.

De Groot, A.S., Sbai, H., Aubin, C.S., McMurry, J. and Martin, W. (2002). Immuno-informatics: Mining genomes for vaccine components, Immunol Cell Biol, *80*, 255-69.

Delves, P.J., Lund, T. and Roitt, I.M. (1997). Can epitope-focused vaccines select advantageous immune responses?, Mol Med Today, *3*, 55-60.

Dertzbaugh, M.T. (1998). Genetically engineered vaccines: an overview, Plasmid, *39*, 100-13.

Donnes, P. and Kohlbacher, O. (2005). Integrated modeling of the major events in the MHC class I antigen processing pathway, Protein Sci, *14*, 2132-40.

Doolan, D.L., Hoffman, S.L., Southwood, S., Wentworth, P.A., Sidney, J., Chesnut, R.W., Keogh, E., Appella, E., Nutman, T.B., Lal, A.A., Gordon, D.M., Oloo, A. and Sette, A. (1997). Degenerate cytotoxic T cell epitopes from P. falciparum restricted by multiple HLA-A and HLA-B supertype alleles, Immunity, *7*, 97-112.

Doytchinova, I.A. and Flower, D.R. (2005). In silico identification of supertypes for class II MHCs, J Immunol, *174*, 7085-95.

Doytchinova, I.A., Guan, P. and Flower, D.R. (2004). Quantitative structure-activity relationships and the prediction of MHC supermotifs, Methods, *34*, 444-53.

Duffy, P.E., Krzych, U., Francis, S. and Fried, M. (2005). Malaria vaccines: using models of immunity and functional genomics tools to accelerate the development of vaccines against Plasmodium falciparum, Vaccine, *23*, 2235-42.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res, *32*, 1792-7.

Egloff, M.P., Benarroch, D., Selisko, B., Romette, J.L. and Canard, B. (2002). An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization, Embo J, *21*, 2757-68.

Endy, T.P., Nisalak, A., Chunsuttitwat, S., Vaughn, D.W., Green, S., Ennis, F.A., Rothman, A.L. and Libraty, D.H. (2004). Relationship of preexisting dengue virus (DV) neutralizing antibody levels to viremia and severity of disease in a prospective cohort study of DV infection in Thailand, J Infect Dis, *189*, 990-1000.

Erbel, P., Schiering, N., D'Arcy, A., Renatus, M., Kroemer, M., Lim, S.P., Yin, Z., Keller, T.H., Vasudevan, S.G. and Hommel, U. (2006). Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus, Nat Struct Mol Biol, *13*, 372-3.

Esser, M.T., Marchese, R.D., Kierstead, L.S., Tussey, L.G., Wang, F., Chirmule, N. and Washabaugh, M.W. (2003). Memory T cells and vaccines, Vaccine, *21*, 419-30.

Evavold, B.D., Sloan-Lancaster, J. and Allen, P.M. (1993). Tickling the TCR: selective T-cell functions stimulated by altered peptide ligands, Immunol Today, *14*, 602-9.

Falgout, B., Chanock, R. and Lai, C.J. (1989). Proper processing of dengue virus nonstructural glycoprotein NS1 requires the N-terminal hydrophobic signal sequence and the downstream nonstructural protein NS2a, J Virol, *63*, 1852-60.

Feito, M.J., Gomez-Gutierrez, J., Ayora, S., Alonso, J.C., Peterson, D. and Gavilanes, F. (2008). Insights into the oligomerization state-helicase activity relationship of West Nile virus NS3 NTPase/helicase, Virus Res, *135*, 166-74.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, *5*, 164-166.

Fischer, W., Perkins, S., Theiler, J., Bhattacharya, T., Yusim, K., Funkhouser, R., Kuiken, C., Haynes, B., Letvin, N.L., Walker, B.D., Hahn, B.H. and Korber, B.T. (2007). Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants, Nat Med, *13*, 100-6.

Freeman, S. (2003) *Biological Science*. Prentice Hall, NJ, USA.

Fu, J., Tan, B.H., Yap, E.H., Chan, Y.C. and Tan, Y.H. (1992). Full-length cDNA sequence of dengue type 1 virus (Singapore strain S275/90), Virology, *188*, 953-8.

Fugger, L., Michie, S.A., Rulifson, I., Lock, C.B. and McDevitt, G.S. (1994). Expression of HLA-DR4 and human CD4 transgenes in mice determines the variable region beta-chain T-cell repertoire and mediates an HLA-DR-restricted immune response, Proc Natl Acad Sci U S A, *91*, 6151-5.

Gagnon, S.J., Zeng, W., Kurane, I. and Ennis, F.A. (1996). Identification of two epitopes on the dengue 4 virus capsid protein recognized by a serotype-specific and a panel of serotype-cross-reactive human CD4+ cytotoxic T-lymphocyte clones, J Virol, *70*, 141-7.

Gao, F., Korber, B.T., Weaver, E., Liao, H.X., Hahn, B.H. and Haynes, B.F. (2004). Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity, Expert Rev Vaccines, *3*, S161-8.

Garcia-Quintanilla, A. (2007). Overcoming viral escape with vaccines that generate and display antigen diversity in vivo, Virol J, *4*, 125.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B.H., Bhattacharya, T. and Korber, B. (2002). Diversity considerations in HIV-1 vaccine selection, Science, *296*, 2354-60.

Gottesman, M.E. and Weisberg, R.A. (2004). Little lambda, who made thee?, Microbiol Mol Biol Rev, *68*, 796-813.

Grandi, G. (2003). Rational antibacterial vaccine design through genomic technologies, Int J Parasitol, *33*, 615-20.

Gubler, D.J. (2001). Human arbovirus infections worldwide, Ann N Y Acad Sci, *951*, 13-24.

Gupta, V., Tabiin, T.M., Sun, K., Chandrasekaran, A., Anwar, A., Yang, K., Chikhlikar, P., Salmon, J., Brusic, V., Marques, E.T., Kellathur, S.N. and August, T.J. (2006). SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens, Virology, *347*, 127-39.

Halstead, S.B. (1988). Pathogenesis of dengue: challenges to molecular biology, Science, *239*, 476-81.

Halstead, S.B. and Deen, J. (2002). The future of dengue vaccines, Lancet, *360*, 1243-5.

Harcourt, G.C., Garrard, S., Davenport, M.P., Edwards, A. and Phillips, R.E. (1998). HIV-1 variation diminishes CD4 T lymphocyte recognition, J Exp Med, *188*, 1785-93.

Hatch, S., Mathew, A. and Rothman, A. (2008). Dengue vaccine: opportunities and challenges, IDrugs, *11*, 42-5.

Haydon, D.T. and Woolhouse, M.E. (1998). Immune avoidance strategies in RNA viruses: fitness continuums arising from trade-offs between immunogenicity and antigenic variability, J Theor Biol, *193*, 601-12.

Hayes, E.B., Sejvar, J.J., Zaki, S.R., Lanciotti, R.S., Bode, A.V. and Campbell, G.L. (2005). Virology, pathology, and clinical manifestations of West Nile virus disease, Emerg Infect Dis, *11*, 1174-9.

Heiny, A.T., Miotto, O., Srinivasan, K.N., Khan, A.M., Zhang, G.L., Brusic, V., Tan, T.W. and August, J.T. (2007). Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets, PLoS ONE, *2*, e1190.

Henchal, E.A. and Putnak, J.R. (1990). The dengue viruses, Clin Microbiol Rev, *3*, 376-96.

Holmes, E.C. and Burch, S.S. (2000). The causes and consequences of genetic variation in dengue virus, Trends Microbiol, *8*, 74-7.

Holmes, E.C. and Twiddy, S.S. (2003). The origin, emergence and evolutionary genetics of dengue virus, Infect Genet Evol, *3*, 19-28.

Holmes, E.C., Worobey, M. and Rambaut, A. (1999). Phylogenetic evidence for recombination in dengue virus, Mol Biol Evol, *16*, 405-9.

Horga, M.A. and Fine, A. (2001). West Nile virus, Pediatr Infect Dis J, *20*, 801-2.

Hudson, A.W. and Ploegh, H.L. (2002). The cell biology of antigen presentation, Exp Cell Res, *272*, 1-7.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. and Sigrist, C.J. (2006). The PROSITE database, Nucleic Acids Res, *34*, D227-30.

Imrie, A., Meeks, J., Gurary, A., Sukhbataar, M., Kitsutani, P., Effler, P. and Zhao, Z. (2007). Differential functional avidity of dengue virus-specific T-cell clones for variant peptides representing heterologous and previously encountered serotypes, J Virol, *81*, 10081-91.

Innis, B.L. and Eckels, K.H. (2003). Progress in development of a live-attenuated, tetravalent dengue virus vaccine by the United States Army Medical Research and Materiel Command, Am J Trop Med Hyg, *69*, 1-4.

Ito, K., Bian, H.J., Molina, M., Han, J., Magram, J., Saar, E., Belunis, C., Bolin, D.R., Arceo, R., Campbell, R., Falcioni, F., Vidovic, D., Hammer, J. and Nagy, Z.A. (1996). HLA-DR4-IE chimeric class II transgenic, murine class II-deficient mice are susceptible to experimental allergic encephalomyelitis, J Exp Med, *183*, 2635-44.

Kalergis, A.M. and Nathenson, S.G. (2000). Altered peptide ligand-mediated TCR antagonism can be modulated by a change in a single amino acid residue within the CDR3 beta of an MHC class I-restricted TCR, J Immunol, *165*, 280-5.

Kanai, R., Kar, K., Anthony, K., Gould, L.H., Ledizet, M., Fikrig, E., Marasco, W.A., Koski, R.A. and Modis, Y. (2006). Crystal structure of west nile virus envelope glycoprotein reveals viral surface epitopes, J Virol, *80*, 11000-8.

Kast, W.M., Brandt, R.M., Sidney, J., Drijfhout, J.W., Kubo, R.T., Grey, H.M., Melief, C.J. and Sette, A. (1994). Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins, J Immunol, *152*, 3904-12.

Khan, A.M. (2005). Mapping targets of immune responses in complete dengue viral genomes. Biochemistry. Singapore, Singapore. Master of Science**:** 1-135.

Khan, A.M., Heiny, A.T., Lee, K.X., Srinivasan, K.N., Tan, T.W., August, J.T. and Brusic, V. (2006a). Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus, BMC Bioinformatics, *7 Suppl 5*, S4.

Khan, A.M., Miotto, O., Heiny, A.T., Salmon, J., Srinivasan, K.N., Nascimento, E.J., Marques, E.T., Jr., Brusic, V., Tan, T.W. and August, J.T. (2006b). A systematic bioinformatics approach for selection of epitope-based vaccine targets, Cell Immunol, *244*, 141-7.

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

Kim, S.K. and DeMars, R. (2001). Epitope clusters in the major outer membrane protein of Chlamydia trachomatis, Curr Opin Immunol, *13*, 429-36.

Klenerman, P., Wu, Y. and Phillips, R. (2002). HIV: current opinion in escapology, Curr Opin Microbiol, *5*, 408-13.

Kulkarni-Kale, U., Bhosle, S. and Kolaskar, A.S. (2005). CEP: a conformational epitope prediction server, Nucleic Acids Res, *33*, W168-71.

Kuno, G., Chang, G.J., Tsuchiya, K.R., Karabatsos, N. and Cropp, C.B. (1998). Phylogeny of the genus Flavivirus, J Virol, *72*, 73-83.

Kurane, I., Dai, L.C., Livingston, P.G., Reed, E. and Ennis, F.A. (1993). Definition of an HLA-DPw2-restricted epitope on NS3, recognized by a dengue virus serotype-cross-reactive human CD4+ CD8- cytotoxic T-cell clone, J Virol, *67*, 6285-8.

Kurane, I., Innis, B.L., Nimmannitya, S., Nisalak, A., Rothman, A.L., Livingston, P.G., Janus, J. and Ennis, F.A. (1990). Human immune responses to dengue viruses, Southeast Asian J Trop Med Public Health, *21*, 658-62.

Kurane, I., Okamoto, Y., Dai, L.C., Zeng, L.L., Brinton, M.A. and Ennis, F.A. (1995). Flavivirus-cross-reactive, HLA-DR15-restricted epitope on NS3 recognized by human CD4+ CD8- cytotoxic T lymphocyte clones, J Gen Virol, *76 (Pt 9)*, 2243-9.

Kurane, I., Zeng, L., Brinton, M.A. and Ennis, F.A. (1998). Definition of an epitope on NS3 recognized by human CD4+ cytotoxic T lymphocyte clones cross-reactive for dengue virus types 2, 3, and 4, Virology, *240*, 169-74.

Kutubuddin, M., Kolaskar, A.S., Galande, S., Gore, M.M., Ghosh, S.N. and Banerjee, K. (1991). Recognition of helper T cell epitopes in envelope (E) glycoprotein of Japanese encephalitis, west Nile and Dengue viruses, Mol Immunol, *28*, 149-54.

Kutzler, M.A. and Weiner, D.B. (2008). DNA vaccines: ready for prime time?, Nat Rev Genet, *9*, 776-88.

Lanciotti, R.S., Roehrig, J.T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K.E., Crabtree, M.B., Scherret, J.H., Hall, R.A., MacKenzie, J.S., Cropp, C.B., Panigrahy, B., Ostlund, E., Schmitt, B., Malkinson, M., Banet, C., Weissman, J., Komar, N., Savage, H.M., Stone, W., McNamara, T. and Gubler, D.J. (1999). Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States, Science, *286*, 2333-7.

Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O. and Nielsen, M. (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions, Eur J Immunol, *35*, 2295-303.

Lauemoller, S.L., Holm, A., Hilden, J., Brunak, S., Holst Nissen, M., Stryhn, A., Ostergaard Pedersen, L. and Buus, S. (2001). Quantitative predictions of peptide binding to MHC class I molecules using specificity matrices and anchor-stratified calibrations, Tissue Antigens, *57*, 405-14.

Leung, D., Schroder, K., White, H., Fang, N.X., Stoermer, M.J., Abbenante, G., Martin, J.L., Young, P.R. and Fairlie, D.P. (2001). Activity of recombinant dengue 2 virus NS3 protease in the presence of a truncated NS2B co-factor, small peptide substrates, and inhibitors, J Biol Chem, *276*, 45762-71.

Leyssen, P., De Clercq, E. and Neyts, J. (2000). Perspectives for the treatment of infections with Flaviviridae, Clin Microbiol Rev, *13*, 67-82, table of contents.

Li, H., Clum, S., You, S., Ebner, K.E. and Padmanabhan, R. (1999). The serine protease and RNA-stimulated nucleoside triphosphatase and RNA helicase functional domains of dengue virus type 2 NS3 converge within a region of 20 amino acids, J Virol, *73*, 3108-16.

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

Lindenbach, B.D. and Rice, C.M. (2003). Molecular biology of flaviviruses, Adv Virus Res, *59*, 23-61.

Livingston, P.G., Kurane, I., Dai, L.C., Okamoto, Y., Lai, C.J., Men, R., Karaki, S., Takiguchi, M. and Ennis, F.A. (1995). Dengue virus-specific, HLA-B35-restricted, human CD8+ cytotoxic T lymphocyte (CTL) clones. Recognition of NS3 amino acids 500 to 508 by CTL clones of two different serotype specificities, J Immunol, *154*, 1287-95.

Locher, C.P., Heinrichs, V., Apt, D. and Whalen, R.G. (2004). Overcoming antigenic diversity and improving vaccines using DNA shuffling and screening technologies, Expert Opin Biol Ther, *4*, 589-97.

Loke, H., Bethell, D.B., Phuong, C.X., Dung, M., Schneider, J., White, N.J., Day, N.P., Farrar, J. and Hill, A.V. (2001). Strong HLA class I--restricted T cell responses in dengue hemorrhagic fever: a double-edged sword?, J Infect Dis, *184*, 1369-73.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A.G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. and Brunak, S. (2004). Definition of supertypes for HLA molecules using clustering of specificity matrices, Immunogenetics, *55*, 797-810.

MacDonald, K.S., Matukas, L., Embree, J.E., Fowke, K., Kimani, J., Nagelkerke, N.J., Oyugi, J., Kiama, P., Kaul, R., Luscher, M.A., Rowland-Jones, S., Ndinya-Achola, J., Ngugi, E., Bwayo, J.J. and Plummer, F.A. (2001). Human leucocyte antigen supertypes and immune susceptibility to HIV-1, implications for vaccine design, Immunol Lett, *79*, 151-7.

Mackenzie, J.M., Kenney, M.T. and Westaway, E.G. (2007). West Nile virus strain Kunjin NS5 polymerase is a phosphoprotein localized at the cytoplasmic site of viral RNA synthesis, J Gen Virol, *88*, 1163-8.

Madrenas, J. and Germain, R.N. (1996). Variant TCR ligands: new insights into the molecular basis of antigen-dependent signal transduction and T-cell activation, Semin Immunol, *8*, 83-101.

Madsen, L., Labrecque, N., Engberg, J., Dierich, A., Svejgaard, A., Benoist, C., Mathis, D. and Fugger, L. (1999). Mice lacking all conventional MHC class II genes, Proc Natl Acad Sci U S A, *96*, 10338-43.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., Tettelin, H., Brettoni, C., Iacobini, E.T., Rosini, R., D'Agostino, N., Miorin, L., Buccato, S., Mariani, M., Galli, G., Nogarotto, R., Nardi Dei, V., Vegni, F., Fraser, C., Mancuso, G., Teti, G., Madoff, L.C., Paoletti, L.C., Rappuoli, R., Kasper, D.L., Telford, J.L. and Grandi, G. (2005). Identification of a universal Group B streptococcus vaccine by multiple genome screen, Science, *309*, 148-50.

Malet, H., Egloff, M.P., Selisko, B., Butcher, R.E., Wright, P.J., Roberts, M., Gruez, A., Sulzenbacher, G., Vonrhein, C., Bricogne, G., Mackenzie, J.M., Khromykh, A.A., Davidson, A.D. and Canard, B. (2007). Crystal structure of the RNA polymerase domain of the West Nile virus non-structural protein 5, J Biol Chem, *282*, 10678-89.

Mangada, M.M. and Rothman, A.L. (2005). Altered cytokine responses of dengue-specific CD4+ T cells to heterologous serotypes, J Immunol, *175*, 2676-83.

Marfin, A.A., Petersen, L.R., Eidson, M., Miller, J., Hadler, J., Farello, C., Werner, B., Campbell, G.L., Layton, M., Smith, P., Bresnitz, E., Cartter, M., Scaletta, J., Obiri, G., Bunning, M., Craven, R.C., Roehrig, J.T., Julian, K.G., Hinten, S.R. and Gubler, D.J. (2001). Widespread West Nile virus activity, eastern United States, 2000, Emerg Infect Dis, *7*, 730-5.

Markoff, L., Falgout, B. and Chang, A. (1997). A conserved internal hydrophobic domain mediates the stable membrane integration of the dengue virus capsid protein, Virology, *233*, 105-17.

Mathew, A., Kurane, I., Rothman, A.L., Zeng, L.L., Brinton, M.A. and Ennis, F.A. (1996). Dominant recognition by human CD8+ cytotoxic T lymphocytes of dengue virus nonstructural proteins NS3 and NS1.2a, J Clin Invest, *98*, 1684-91.

Mazumder, R., Hu, Z.Z., Vinayaka, C.R., Sagripanti, J.L., Frost, S.D., Kosakovsky Pond, S.L. and Wu, C.H. (2007). Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines, Virus Genes, *35*, 175-86.

McGinnis, S. and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res, *32*, W20-5.

Miotto, O., Heiny, A., Tan, T.W., August, J.T. and Brusic, V. (2008). Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis, BMC Bioinformatics, *9 Suppl 1*, S18.

Miotto, O., Tan, T.W. and Brusic, V. (2005) Extraction by Example: Induction of Structural Rules for the Analysis of Molecular Sequence Data from

Heterogeneous Sources. In M. Gallagher, J. Hogan and F. Maire (eds), *Lecture Notes in Computer Science 3578*. Springer, Berlin, 398-405.

Modis, Y., Ogata, S., Clements, D. and Harrison, S.C. (2004). Structure of the dengue virus envelope protein after membrane fusion, Nature, *427*, 313-9.

Monath, T.P. (1994). Dengue: the risk to developed and developing countries, Proc Natl Acad Sci U S A, *91*, 2395-400.

Mongkolsapaya, J., Dejnirattisai, W., Xu, X.N., Vasanawathana, S., Tangthawornchaikul, N., Chairunsri, A., Sawasdivorn, S., Duangchinda, T., Dong, T., Rowland-Jones, S., Yenchitsomanus, P.T., McMichael, A., Malasit, P. and Screaton, G. (2003). Original antigenic sin and apoptosis in the pathogenesis of dengue hemorrhagic fever, Nat Med, *9*, 921-7.

Mongkolsapaya, J., Duangchinda, T., Dejnirattisai, W., Vasanawathana, S., Avirutnan, P., Jairungsri, A., Khemnu, N., Tangthawornchaikul, N., Chotiyarnwong, P., Sae-Jang, K., Koch, M., Jones, Y., McMichael, A., Xu, X., Malasit, P. and Screaton, G. (2006). T cell responses in dengue hemorrhagic fever: are cross-reactive T cells suboptimal?, J Immunol, *176*, 3821-9.

Moran, E., Simmons, C., Vinh Chau, N., Luhn, K., Wills, B., Dung, N.P., Thao le, T.T., Hien, T.T., Farrar, J., Rowland-Jones, S. and Dong, T. (2008). Preservation of a critical epitope core region is associated with the high degree of flaviviral cross-reactivity exhibited by a dengue-specific CD4(+) T cell clone, Eur J Immunol, *38*, 1050-7.

Morvan, J., Besselaar, T., Fontenille, D. and Coulanges, P. (1990). Antigenic variations in West Nile virus strains isolated in Madagascar since 1978, Res Virol, *141*, 667-76.

Mukhopadhyay, S., Kim, B.S., Chipman, P.R., Rossmann, M.G. and Kuhn, R.J. (2003). Structure of West Nile virus, Science, *302*, 248.

Mukhopadhyay, S., Kuhn, R.J. and Rossmann, M.G. (2005). A structural perspective of the flavivirus life cycle, Nat Rev Microbiol, *3*, 13-22.

Murthy, H.M., Clum, S. and Padmanabhan, R. (1999). Dengue virus NS3 serine protease. Crystal structure and insights into interaction of the active site with substrates by molecular modeling and structural analysis of mutational effects, J Biol Chem, *274*, 5573-80.

Muzzi, A., Masignani, V. and Rappuoli, R. (2007). The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials, Drug Discov Today, *12*, 429-39.

Nishimura, Y., Chen, Y.Z., Uemura, Y., Tanaka, Y., Tsukamoto, H., Kanai, T., Yokomizo, H., Yun, C., Matsuoka, T., Irie, A. and Matsushita, S. (2004).

Degenerate recognition and response of human CD4+ Th cell clones: implications for basic and applied immunology, Mol Immunol, *40*, 1089-94.

Novitsky, V., Smith, U.R., Gilbert, P., McLane, M.F., Chigwedere, P., Williamson, C., Ndung'u, T., Klein, I., Chang, S.Y., Peter, T., Thior, I., Foley, B.T., Gaolekwe, S., Rybak, N., Gaseitsiwe, S., Vannberg, F., Marlink, R., Lee, T.H. and Essex, M. (2002). Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design?, J Virol, *76*, 5435-51.

Okamoto, Y., Kurane, I., Leporati, A.M. and Ennis, F.A. (1998). Definition of the region on NS3 which contains multiple epitopes recognized by dengue virus serotype-cross-reactive and flavivirus-cross-reactive, HLA-DPw2-restricted CD4+ T cell clones, J Gen Virol, *79 (Pt 4)*, 697-704.

Osatomi, K. and Sumiyoshi, H. (1990). Complete nucleotide sequence of dengue type 3 virus genome RNA, Virology, *176*, 643-7.

Ovsyannikova, I.G., Jacobson, R.M. and Poland, G.A. (2004). Variation in vaccine response in normal populations, Pharmacogenomics, *5*, 417-27.

Page, R.D. (2002). Visualizing phylogenetic trees using TreeView, Curr Protoc Bioinformatics, *Chapter 6*, Unit 6 2.

Paninski, L. (2003). Estimation of entropy and mutual information, Neural Computation, *15*, 1191-1253.

Pappalardo, F., Halling-Brown, M.D., Rapin, N., Zhang, P., Alemani, D., Emerson, A., Paci, P., Duroux, P., Pennisi, M., Palladini, A., Miotto, O., Churchill, D., Rossi, E., Shepherd, A.J., Moss, D.S., Castiglione, F., Bernaschi, M., Lefranc, M.P., Brunak, S., Motta, S., Lollini, P.L., Basford, K.E. and Brusic, V. (2009). ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization, Brief Bioinform, *10*, 330-40.

Parra, M., Hui, G., Johnson, A.H., Berzofsky, J.A., Roberts, T., Quakyi, I.A. and Taylor, D.W. (2000). Characterization of conserved T- and B-cell epitopes in Plasmodium falciparum major merozoite surface protein 1, Infect Immun, *68*, 2685-91.

Peters, B., Sidney, J., Bourne, P., Bui, H.H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo, R., Lund, O., Nemazee, D., Ponomarenko, J.V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Wilson, S. and Sette, A. (2005). The immune epitope database and analysis resource: from vision to blueprint, PLoS Biol, *3*, e91.

Petersen, L.R., Marfin, A.A. and Gubler, D.J. (2003). West Nile virus, JAMA, *290*, 524-8.

Petersen, L.R. and Roehrig, J.T. (2001). West Nile virus: a reemerging global pathogen, Emerg Infect Dis, *7*, 611-4.

Petrovsky, N. and Aguilar, J.C. (2004). Vaccine adjuvants: current state and future trends, Immunol Cell Biol, *82*, 488-96.

Petrovsky, N., Silva, D. and Brusic, V. (2003). The future for computational modelling and prediction systems in clinical immunology, Novartis Found Symp, *254*, 23-32; discussion 33-42, 98-101, 250-2.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D.W., Jeffries, A.C., Saunders, N.J., Granoff, D.M., Venter, J.C., Moxon, E.R., Grandi, G. and Rappuoli, R. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing, Science, *287*, 1816-20.

Poland, G.A., Ovsyannikova, I.G. and Jacobson, R.M. (2008). Personalized vaccines: the emerging field of vaccinomics, Expert Opin Biol Ther, *8*, 1659-67.

Preugschat, F. and Strauss, J.H. (1991). Processing of nonstructural proteins NS4A and NS4B of dengue 2 virus in vitro and in vivo, Virology, *185*, 689-97.

Price, G.E., Ou, R., Jiang, H., Huang, L. and Moskophidis, D. (2000). Viral escape by selection of cytotoxic T cell-resistant variants in influenza A virus pneumonia, J Exp Med, *191*, 1853-67.

Pulendran, B. and Ahmed, R. (2006). Translating innate immunity into immunological memory: implications for vaccine development, Cell, *124*, 849-63.

Rammensee, H.G. (1995). Chemistry of peptides associated with MHC class I and class II molecules, Curr Opin Immunol, *7*, 85-96.

Rammensee, H.G., Friede, T. and Stevanoviic, S. (1995). MHC ligands and peptide motifs: first listing, Immunogenetics, *41*, 178-228.

Rappuoli, R. (2000). Reverse vaccinology, Curr Opin Microbiol, *3*, 445-50.

Rappuoli, R. and Covacci, A. (2003). Reverse vaccinology and genomics, Science, *302*, 602.

Rico-Hesse, R. (1990). Molecular evolution and distribution of dengue viruses type 1 and 2 in nature, Virology, *174*, 479-93.

Rico-Hesse, R. (2003). Microevolution and virulence of dengue viruses, Adv Virus Res, *59*, 315-41.

Rimmelzwaan, G.F., Boon, A.C., Voeten, J.T., Berkhoff, E.G., Fouchier, R.A. and Osterhaus, A.D. (2004). Sequence variation in the influenza A virus nucleoprotein associated with escape from cytotoxic T lymphocytes, Virus Res, *103*, 97-100.

Robinson, H.L. and Amara, R.R. (2005). T cell vaccines for microbial infections, Nat Med, *11*, S25-32.

Roederer, M. and Koup, R.A. (2003). Optimized determination of T cell epitope responses, J Immunol Methods, *274*, 221-8.

Roehrig, J.T. (2003). Antigenic structure of flavivirus proteins, Adv Virus Res, *59*, 141-75.

Rosloniec, E.F., Brand, D.D., Myers, L.K., Whittington, K.B., Gumanovskaya, M., Zaller, D.M., Woods, A., Altmann, D.M., Stuart, J.M. and Kang, A.H. (1997). An HLA-DR1 transgene confers susceptibility to collagen-induced arthritis elicited with human type II collagen, J Exp Med, *185*, 1113-22.

Rothman, A.L. (2004). Dengue: defining protective versus pathologic immunity, J Clin Invest, *113*, 946-51.

Rothman, A.L., Kurane, I., Zhang, Y.M., Lai, C.J. and Ennis, F.A. (1989). Dengue virus-specific murine T-lymphocyte proliferation: serotype specificity and response to recombinant viral proteins, J Virol, *63*, 2486-91.

Sanchez, V., Gimenez, S., Tomlinson, B., Chan, P.K., Thomas, G.N., Forrat, R., Chambonneau, L., Deauvieau, F., Lang, J. and Guy, B. (2006). Innate and adaptive cellular immunity in flavivirus-naive human recipients of a live-attenuated dengue serotype 3 vaccine produced in Vero cells (VDV3), Vaccine, *24*, 4914-26.

Schein, C.H., Zhou, B. and Braun, W. (2005). Stereophysicochemical variability plots highlight conserved antigenic areas in Flaviviruses, Virol J, *2*, 40.

Screaton, G. and Mongkolsapaya, J. (2006). T cell responses and dengue haemorrhagic fever, Novartis Found Symp, *277*, 164-71; discussion 171-6, 251-3.

Sette, A. and Fikes, J. (2003). Epitope-based vaccines: an update on epitope identification, vaccine design and delivery, Curr Opin Immunol, *15*, 461-70.

Sette, A., Livingston, B., McKinney, D., Appella, E., Fikes, J., Sidney, J., Newman, M. and Chesnut, R. (2001). The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation, Biologicals, *29*, 271-6.

Sette, A. and Sidney, J. (1998). HLA supertypes and supermotifs: a functional perspective on HLA polymorphism, Curr Opin Immunol, *10*, 478-82.

Sette, A. and Sidney, J. (1999). Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism, Immunogenetics, *50*, 201-12.

Sette, A., Sidney, J., Livingston, B.D., Dzuris, J.L., Crimi, C., Walker, C.M., Southwood, S., Collins, E.J. and Hughes, A.L. (2003). Class I molecules with similar peptide-binding specificities are the result of both common ancestry and convergent evolution, Immunogenetics, *54*, 830-41.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W.M., Melief, C.J., Oseroff, C., Yuan, L., Ruppert, J. and et al. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes, J Immunol, *153*, 5586-92.

Shankar, P., Fabry, J.A., Fong, D.M. and Lieberman, J. (1996). Three regions of HIV-1 gp160 contain clusters of immunodominant CTL epitopes, Immunol Lett, *52*, 23-30.

Shannon, C.E. (1948). A mathematical theory of communication, Bell System Technical Journal, *27*, 379-423 and 623-656.

Shastri, N., Schwab, S. and Serwold, T. (2002). Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules, Annu Rev Immunol, *20*, 463-93.

Sidney, J., Grey, H.M., Kubo, R.T. and Sette, A. (1996). Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs, Immunol Today, *17*, 261-6.

Sidney, J., Peters, B., Frahm, N., Brander, C. and Sette, A. (2008). HLA class I supertypes: a revised and updated classification, BMC Immunol, *9*, 1.

Sidney, J., Southwood, S., Oseroff, C., del Guercio, M.F., Grey, H.M. and Sette, A. (1998). Measurement of MHC/peptide interactions by gel filtration., Current Protocols in Immunology, 18.3.1-18.3.19.

Simmons, C.P., Dong, T., Chau, N.V., Dung, N.T., Chau, T.N., Thao le, T.T., Dung, N.T., Hien, T.T., Rowland-Jones, S. and Farrar, J. (2005). Early T-cell responses to dengue virus epitopes in Vietnamese adults with secondary dengue virus infections, J Virol, *79*, 5665-75.

Sloan-Lancaster, J. and Allen, P.M. (1996). Altered peptide ligand-induced partial T cell activation: molecular mechanisms and role in T cell biology, Annu Rev Immunol, *14*, 1-27.

Slobod, K.S., Bonsignori, M., Brown, S.A., Zhan, X., Stambas, J. and Hurwitz, J.L. (2005). HIV vaccines: brief review and discussion of future directions, Expert Rev Vaccines, *4*, 305-13.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D. and Fouchier, R.A. (2004). Mapping the antigenic and genetic evolution of influenza virus, Science, *305*, 371-6.

Smithburn, K.C., Hughes, T.P., Burke, A.W. and Paul, J.H. (1940). A neurotropic virus isolated from the blood of a native of Uganda, American Journal of Tropical Medicine, *20*, 471–492.

Solomon, T. and Mallewa, M. (2001). Dengue and other emerging flaviviruses, J Infect, *42*, 104-15.

Southwood, S., Sidney, J., Kondo, A., del Guercio, M.F., Appella, E., Hoffman, S., Kubo, R.T., Chesnut, R.W., Grey, H.M. and Sette, A. (1998). Several common HLA-DR types share largely overlapping peptide binding repertoires, J Immunol, *160*, 3363-73.

Srinivasan, K.N., Gopalakrishnakone, P., Tan, P.T., Chew, K.C., Cheng, B., Kini, R.M., Koh, J.L., Seah, S.H. and Brusic, V. (2002). SCORPION, a molecular database of scorpion toxins, Toxicon, *40*, 23-31.

Srinivasan, K.N., Zhang, G.L., Khan, A.M., August, J.T. and Brusic, V. (2004). Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens, Bioinformatics, *20 Suppl 1*, I297-I302.

Stenman, U.H. (2001). Immunoassay standardization: is it possible, who is responsible, who is capable?, Clin Chem, *47*, 815-20.

Strauss, G., Vignali, D.A., Schonrich, G. and Hammerling, G.J. (1994). Negative and positive selection by HLA-DR3(DRw17) molecules in transgenic mice, Immunogenetics, *40*, 104-8.

Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F. and Hammer, J. (1999). Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices, Nat Biotechnol, *17*, 555-61.

Surman, S., Lockey, T.D., Slobod, K.S., Jones, B., Riberdy, J.M., White, S.W., Doherty, P.C. and Hurwitz, J.L. (2001). Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing, Proc Natl Acad Sci U S A, *98*, 4587-92.

Sylvester-Hvid, C., Kristensen, N., Blicher, T., Ferre, H., Lauemoller, S.L., Wolf, X.A., Lamberth, K., Nissen, M.H., Pedersen, L.O. and Buus, S. (2002). Establishment of a quantitative ELISA capable of determining peptide - MHC class I interaction, Tissue Antigens, *59*, 251-8.

Takahashi, H., Merli, S., Putney, S.D., Houghten, R., Moss, B., Germain, R.N. and Berzofsky, J.A. (1989). A single amino acid interchange yields reciprocal CTL specificities for HIV-1 gp160, Science, *246*, 118-21.

Thomas, P.G., Keating, R., Hulse-Post, D.J. and Doherty, P.C. (2006). Cell-mediated protection in influenza infection, Emerg Infect Dis, *12*, 48-54.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res, *25*, 4876-82.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res, *22*, 4673-80.

Thomson, S.A., Jaramillo, A.B., Shoobridge, M., Dunstan, K.J., Everett, B., Ranasinghe, C., Kent, S.J., Gao, K., Medveckzy, J., Ffrench, R.A. and Ramshaw, I.A. (2005). Development of a synthetic consensus sequence scrambled antigen HIV-1 vaccine designed for global use, Vaccine, *23*, 4647-57.

Tishon, A., LaFace, D.M., Lewicki, H., van Binnendijk, R.S., Osterhaus, A. and Oldstone, M.B. (2000). Transgenic mice expressing human HLA and CD8 molecules generate HLA-restricted measles virus cytotoxic T lymphocytes of the same specificity as humans with natural measles virus infection, Virology, *275*, 286-93.

Tolou, H.J., Couissinier-Paris, P., Durand, J.P., Mercier, V., de Pina, J.J., de Micco, P., Billoir, F., Charrel, R.N. and de Lamballerie, X. (2001). Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences, J Gen Virol, *82*, 1283-90.

Tong, J.C., Zhang, G.L., Tan, T.W., August, J.T., Brusic, V. and Ranganathan, S. (2006). Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides, Bioinformatics, *22*, 1232-8.

Trachtenberg, E., Korber, B., Sollars, C., Kepler, T.B., Hraber, P.T., Hayes, E., Funkhouser, R., Fugate, M., Theiler, J., Hsu, Y.S., Kunstman, K., Wu, S., Phair, J., Erlich, H. and Wolinsky, S. (2003). Advantage of rare HLA supertype in HIV disease progression, Nat Med, *9*, 928-35.

Trent, D.W., Grant, J.A., Rosen, L. and Monath, T.P. (1983). Genetic variation among dengue 2 viruses of different geographic origin, Virology, *128*, 271-84.

Twiddy, S.S., Farrar, J.J., Vinh Chau, N., Wills, B., Gould, E.A., Gritsun, T., Lloyd, G. and Holmes, E.C. (2002). Phylogenetic relationships and differential selection pressures among genotypes of dengue-2 virus, Virology, *298*, 63-72.

Twiddy, S.S., Holmes, E.C. and Rambaut, A. (2003). Inferring the rate and time-scale of dengue virus evolution, Mol Biol Evol, *20*, 122-9.

Ueno, T., Idegami, Y., Motozono, C., Oka, S. and Takiguchi, M. (2007). Altering effects of antigenic variations in HIV-1 on antiviral effectiveness of HIV-specific CTLs, J Immunol, *178*, 5513-23.

Ulmer, J.B., Valley, U. and Rappuoli, R. (2006). Vaccine manufacturing: challenges and solutions, Nat Biotechnol, *24*, 1377-83.

Uzcategui, N.Y., Camacho, D., Comach, G., Cuello de Uzcategui, R., Holmes, E.C. and Gould, E.A. (2001). Molecular epidemiology of dengue type 2 virus in Venezuela: evidence for in situ virus evolution and recombination, J Gen Virol, *82*, 2945-53.

Valdar, W.S. (2002). Scoring residue conservation, Proteins, *48*, 227-41.

Vandenbark, A.A., Rich, C., Mooney, J., Zamora, A., Wang, C., Huan, J., Fugger, L., Offner, H., Jones, R. and Burrows, G.G. (2003). Recombinant TCR ligand induces tolerance to myelin oligodendrocyte glycoprotein 35-55 peptide and reverses clinical and histological signs of chronic experimental autoimmune encephalomyelitis in HLA-DR2 transgenic mice, J Immunol, *171*, 127-33.

Vazquez, S., Guzman, M.G., Guillen, G., Chinea, G., Perez, A.B., Pupo, M., Rodriguez, R., Reyes, O., Garay, H.E., Delgado, I., Garcia, G. and Alvarez, M. (2002). Immune response to synthetic peptides of dengue prM protein, Vaccine, *20*, 1823-30.

Vernikos, G.S. (2008). Overtake in reverse gear, Nature Reviews Microbiology, *6*, 334-335.

Voeten, J.T., Bestebroer, T.M., Nieuwkoop, N.J., Fouchier, R.A., Osterhaus, A.D. and Rimmelzwaan, G.F. (2000). Antigenic drift in the influenza A virus (H3N2) nucleoprotein and escape from recognition by cytotoxic T lymphocytes, J Virol, *74*, 6800-7.

Wagner, R., Leschonsky, B., Harrer, E., Paulus, C., Weber, C., Walker, B.D., Buchbinder, S., Wolf, H., Kalden, J.R. and Harrer, T. (1999). Molecular and functional analysis of a conserved CTL epitope in HIV-1 p24 recognized from a long-term nonprogressor: constraints on immune escape associated with targeting a sequence essential for viral replication, J Immunol, *162*, 3727-34.

Wallis, T.P., Huang, C.Y., Nimkar, S.B., Young, P.R. and Gorman, J.J. (2004). Determination of the disulfide bond arrangement of dengue virus NS1 protein, J Biol Chem, *279*, 20729-41.

Wang, P., Sidney, J., Dow, C., Mothe, B., Sette, A. and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach, PLoS Comput Biol, *4*, e1000048.

Wang, W.K., Lin, S.R., Lee, C.M., King, C.C. and Chang, S.C. (2002a). Dengue type 3 virus in plasma is a population of closely related genomes: quasispecies, J Virol, *76*, 4662-5.

Wang, W.K., Sung, T.L., Lee, C.N., Lin, T.Y. and King, C.C. (2002b). Sequence diversity of the capsid gene and the nonstructural gene NS2B of dengue-3 virus in vivo, Virology, *303*, 181-91.

Watts, C. and Amigorena, S. (2001). Phagocytosis and antigen presentation, Semin Immunol, *13*, 373-9.

Wedemeyer, H., Mizukoshi, E., Davis, A.R., Bennink, J.R. and Rehermann, B. (2001). Cross-reactivity between hepatitis C virus and Influenza A virus determinant-specific cytotoxic T cells, J Virol, *75*, 11392-400.

Welsh, R.M. and Fujinami, R.S. (2007). Pathogenic epitopes, heterologous immunity and vaccine design, Nat Rev Microbiol, *5*, 555-63.

Welsh, R.M. and Rothman, A.L. (2003). Dengue immune response: low affinity, high febrility, Nat Med, *9*, 820-2.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2005). Database resources of the National Center for Biotechnology Information, Nucleic Acids Res, *33*, D39-45.

Whitehead, S.S., Blaney, J.E., Durbin, A.P. and Murphy, B.R. (2007). Prospects for a dengue virus vaccine, Nat Rev Microbiol, *5*, 518-28.

Wilder-Smith, A. and Deen, J.L. (2008). Dengue vaccines for travelers, Expert Rev Vaccines, *7*, 569-78.

Williams, T.M. (2001). Human leukocyte antigen gene polymorphism and the histocompatibility laboratory, J Mol Diagn, *3*, 98-104.

Wilson, C.C., McKinney, D., Anders, M., MaWhinney, S., Forster, J., Crimi, C., Southwood, S., Sette, A., Chesnut, R., Newman, M.J. and Livingston, B.D. (2003). Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1, J Immunol, *171*, 5611-23.

Worobey, M., Rambaut, A. and Holmes, E.C. (1999). Widespread intra-serotype recombination in natural populations of dengue virus, Proc Natl Acad Sci U S A, *96*, 7352-7.

Xu, T., Sampath, A., Chao, A., Wen, D., Nanao, M., Chene, P., Vasudevan, S.G. and Lescar, J. (2005). Structure of the Dengue virus helicase/nucleoside triphosphatase catalytic domain at a resolution of 2.4 A, J Virol, *79*, 10278-88.

Yap, T.L., Xu, T., Chen, Y.L., Malet, H., Egloff, M.P., Canard, B., Vasudevan, S.G. and Lescar, J. (2007). Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution, J Virol, *81*, 4753-65.

Zeng, L., Kurane, I., Okamoto, Y., Ennis, F.A. and Brinton, M.A. (1996). Identification of amino acids involved in recognition by dengue virus NS3-specific, HLA-DR15-restricted cytotoxic CD4+ T-cell clones, J Virol, *70*, 3108-17.

Zhang, C., Mammen, M.P., Jr., Chinnawirotpisan, P., Klungthong, C., Rodpradit, P., Monkongdee, P., Nimmannitya, S., Kalayanarooj, S. and Holmes, E.C. (2005a). Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence, J Virol, *79*, 15123-30.

Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T. and Brusic, V. (2005b). MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides, Nucleic Acids Res, *33*, W172-9.

Zhang, G.L., Khan, A.M., Srinivasan, K.N., Heiny, A., Lee, K., Kwoh, C.K., August, J.T. and Brusic, V. (2008). Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes, BMC Bioinformatics, *9 Suppl 1*, S19.

Zhang, G.L., Petrovsky, N., Kwoh, C.K., August, J.T. and Brusic, V. (2006). PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing, Immunome Res, *2*, 3.

Zhang, L., Mohan, P.M. and Padmanabhan, R. (1992). Processing and localization of Dengue virus type 2 polyprotein precursor NS3-NS4A-NS4B-NS5, J Virol, *66*, 7549-54.

Zinkernagel, R.M. and Hengartner, H. (2004). On immunity against infections and vaccines: credo 2004, Scand J Immunol, *60*, 9-13.

Zivny, J., DeFronzo, M., Jarry, W., Jameson, J., Cruz, J., Ennis, F.A. and Rothman, A.L. (1999). Partial agonist effect influences the CTL response to a heterologous dengue virus serotype, J Immunol, *163*, 2754-60.

Zulueta, A., Martin, J., Hermida, L., Alvarez, M., Valdes, I., Prado, I., Chinea, G., Rosario, D., Guillen, G. and Guzman, M.G. (2006). Amino acid changes in the recombinant Dengue 3 Envelope domain III determine its antigenicity and immunogenicity in mice, Virus Res, *121*, 65-73.

Author’s Publications

PUBLICATIONS 2003-2009

Related to the thesis

Journal articles

1. Koo QY, **Khan AM**, Jung KO, Ramdas S, Miotto O, Tan TW, Brusic V, Salmon J, August JT (2009) Conservation and variability of West Nile virus proteins. PLoS ONE 4(4): e5352. [**IF 2008 = N.A. (new); Citations: N.A**] (*contributed equally as the first author*)

2. **Khan AM**, Miotto O, Nascimento EJ, Srinivasan KN, Heiny AT, Zhang GL, Marques ET, Tan TW, Brusic V, Salmon J, August JT (2008) Conservation and variability of dengue virus proteins: implications for vaccine design. PLoS Negl. Trop. Dis. 2 (8), e272. PMID: 18698358 [**IF 2008 = 4.172; Citations: 8**]

3. **Khan AM**, Miotto O, Heiny AT, Salmon J, Srinivasan KN, Nascimento EJ, Marques ET, Brusic V, Tan TW, August JT (2006) A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cell Immunol. 244(2), 141-7. PMID: 17434154 **[IF 2008 = 1.893; Citations = 15]**

4. **Khan AM**, Heiny AT, Lee KX, Srinivasan KN, Tan TW, August JT, Brusic V (2006) Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. BMC Bioinformatics. 7 Suppl 5, S4. PMID: 17254309 **[IF 2008 = 3.781; Citations = 14]**

Others

Journal articles

5. Mercier PL, Acheson NH, Fauquet CM, **Khan AM**, Lefkowitz EJ, Tordo N (2009) Proposal for a global virus isolate nomenclature. (*Manuscript in preparation*).

6. Jung KO, **Khan AM**, Hu Y, Tan YLB, Simon G, Nascimento EJN, Brusic V, Marques ETA, Tan TW, Salmon J, August JT (2009) West Nile Virus HLA-restricted class I and II T-cell epitope peptides identified by immunizaton of HLA transgenic mice and peptide-specific T-cell activation. (*Manuscript submitted to PLoS Neglected Tropical Diseases; contributed equally as the first author)*.

7. Simon G, **Khan AM**, Zhou J, Salmon J, Chikhlikar PR, Jung KO, Marques ETA, August JT (2009) Efficient activation of HLA DR4-restricted T-cell epitope responses by dendritic cell mediated delivery of plasmid DNA encoding LAMP/HIV-1 Gag immunogen and analysis of epitope peptide variability. PLoS One. 2010 Jan 5;5(1):e8574. [**IF 2008 = N.A. (new); Citations: N.A**]

8. Zhang GL, **Khan AM**, Srinivasan KN, Heiny A, Lee K, Kwoh CK, August JT, Brusic V (2008) Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. BMC Bioinformatics. 9 Suppl 1, S19. PMID: 18315850 **[IF 2008 = 3.781; Citations: 4]** (*contributed equally as the first author*)

9. Sangket U, Phongdara A, Chotigeat W, Nathan D, Kim WY, Bhak J, Ngamphiw C, Tongsima S, **Khan AM**, Lin H, Tan TW (2008) Automatic synchronization and distribution of biological databases and software over low-bandwidth networks among developing countries. Bioinformatics. 24 (2), 299-301. PMID: 18037613 **[IF 2008 = 4.328; Citations = 2]**

10. Heiny AT, Miotto O, Srinivasan KN, **Khan AM**, Zhang GL, Brusic V, Tan TW, August JT (2007) Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. PLoS ONE. 2(11), e1190. PMID: 18030326 **[IF 2008 = N.A (new journal); Citations = 24]**

11. Zhang GL, **Khan AM**, Srinivasan KN, August JT, Brusic V (2005) Neural models for predicting viral vaccine targets. J Bioinform. Comput. Biol. 3 (5), 1207-25. PMID: 16278955 **[IF 2008 = N.A (new journal); Citations = 10]**

12. Zhang GL, **Khan AM**, Srinivasan KN, August JT, Brusic V (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res. 33 (Web Server issue), W172-9. PMID: 15980449 **[IF 2008 = 6.878; Citations = 53]**

13. Siew JP, **Khan AM**, Tan PT, Koh JL, Seah SH, Koo CY, Chai SC, Armugam A, Brusic V, Jeyaseelan K (2004) Systematic analysis of snake neurotoxins' functional classification using a data warehousing approach. Bioinformatics. 20 (18), 3466-80. PMID: 15271784 **[IF 2008 = 4.328; Citations = 6]**

14. Srinivasan KN, Zhang GL, **Khan AM**, August JT, Brusic V (2004) Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. Bioinformatics. 20 Suppl. 1, i297-302. PMID: 15262812 **[IF 2008 = 4.328; Citations = 26]**

15. Lenffer J, Lai P, El Mejaber W, **Khan AM**, Koh JL, Tan PT, Seah SH, Brusic V (2004) CysView: protein classification based on cysteine pairing patterns. Nucleic Acids Res. 32 (Web Server issue), W350-5. PMID: 15215409 **[IF 2008 = 6.878; Citations = 10]**

16. Brahmachary M, Krishnan SP, Koh JL, **Khan AM**, Seah SH, Tan TW, Brusic V, Bajic VB (2004) ANTIMIC: a database of antimicrobial sequences. Nucleic Acids Res. 32 (Database issue), D586-9. PMID: 14681487 **[IF 2008 = 6.878; Citations = 55]**

17. Fry BG, Wüster W, Kini RM, Brusic V, **Khan A**, Venkataraman D, Rooney AP (2003) Molecular evolution and phylogeny of elapid snake venom three-finger toxins. J Mol Evol. 57 (1), 110-29. PMID: 12962311 **[IF 2008 = 2.762; Citations = 59; Award = Zuckerkandl prize]**

18. Tan PT, **Khan AM**, Brusic V (2003) Bioinformatics for venom and toxin sciences. Brief Bioinform. 4 (1), 53-62. PMID: 12715834 **[IF 2008 = 4.627; Citations = 16]**

Conference articles

19. Rajapakse M, Veeramani A, Gopalakrishnan K, Ananthanarayan S, Srinivasan KN, August JT, **Khan AM**, Brusic V (2006) Temporal and antigenic analysis of dengue virus serotype 1 genome polyprotein sequences. International Conference on Biomedical and Pharmaceutical Engineering, 2006. Singapore, December 11, 2006.

20. Koh JLY, Krishnan SPT, Seah SH, Tan PT, **Khan AM**, Lee ML, Brusic V (2004) BioWare: A framework for bioinformatics data retrieval, annotation and publishing. Search and Discovery in Bioinformatics. 27th Annual International ACM SIGIR Conference on Research and Development in IR. Sheffield, UK, July 29, 2004. **[Citations = 9]**

21. Koh JLY, Lee ML, **Khan AM**, Tan PTJ, Brusic V (2004) Duplicate Detection in Biological Data using Association. 2nd European Workshop on Data Mining and Text Mining for Bioinformatics. Pisa, Italy, September 24, 2004. **[Citations = 13]**

Books

22. **Khan AM** (2005) Mapping targets of immune responses in complete dengue viral genomes, Masters Thesis, National University of Singapore, Singapore (186 pages)

23. Brusic V, **Khan AM** (editors) (2005) Abstract book of the 3rd Asia-Pacific Bioinformatics Conference & Singapore Bioinformatics Week, World Scientific, Singapore (142 pages).

24. **Khan AM** (2002) Snake venom PLA2: bioinformatics approach, Honours Thesis, National University of Singapore, Singapore (123 pages).

Appendices

Appendix 1: Catalogue of experimentally mapped DENV T-cell epitopes in humans*

protein	Epitope		Type of cellular immune response¶	HLA restricting molecule	Serotype of infection or of vaccine received¥	Reference (s)
	#	Sequence (amino acid position)				
C	1	DENV-2: PFNMLKRERNRVSTVQQLTK (12-31)	Not available	Not available	Not available	Simmons *et al.*, 2005
	2	DENV-2: RVSTVQQLTKRFSLGMLQGR (22-41)				
	3	DENV-4: KGPLRMVLAFITFLR (41-55) DENV-4: VLAFITFLR (47-55)	$CD4^+$ CTL	DPw4¢	DENV-4	Gagnon *et al.*, 1996
	4	DENV-2: TAGILKRWGTIKKSKAINVL (62-81)	$CD4^+$ (helper/cytotoxic not specified)	Not available	Not available	Simmons *et al.*, 2005
	5	DENV-1: LRGFKKEISNML (81-92)	$CD4^+$ (helper/cytotoxic not indicated)	DPw4	DENV-4	Mangada and Rothman, 2005
	6	DENV-2: LRGFRKEIGRML (81-92)				
	7	DENV-3: LKGFKKEISNML (81-92)				
	8	DENV-4: LIGFRKEIGRML (80-91)	$CD4^+$ (helper/cytotoxic not indicated)	DPw4	DENV-4	Mangada and Rothman, 2005
		DENV-4: LIGFRKEIGRMLNIL (81-95) DENV-4: FRKEIGRML (84-92) DENV-4: FRKEIGRM (84-91) DENV-4: GFRKEIGR (83-90)	$CD4^+$ CTL	DPw4, DR1	DENV-4	Gagnon *et al.*, 1996
preM	1	DENV-2: MSSEGAWKHVQRIETWILRH (20-39)	Not available	Not available	Not available	Simmons *et al.*, 2005
	2	DENV-2: QRIETWILRHPGFTMMAAI (30-49)				
	3	DENV-2: LGELCEDTITYKCPLLRQNE (41-60)	$CD4^+$ (helper/cytotoxic not specified)	Not available	Not available	Simmons *et al.*, 2005

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
E	1	DENV-2: FVEGVSGGSWVDIVL (11-25)	Not available	Not available	Not available	Simmons *et al.*, 2005
	2	DENV-2: SGGSWVDIVLEHGSC (16-30)				
	3	DENV-2: LRKYCIEAKLTNTTT (56-70)				
	4	DENV-2: TLVTFKNPHAKKQDV (136-150)				
	5	DENV-2: VTMECSPRTGLDFNE (181-195)				
	6	DENV-2: MENKAWLVHRQWFLD (201-125)				
	7	DENV-2: KKQDVVVLGSQEGAM (246-260)				
	8	DENV-1: FLDLPLPWT (213-221)	$CD8^+$	A*0201	DENV-1, -2, -3	Bashyam *et al.*, 2006
	9	DENV-2: FLDLPLPWL (213-221)				
	10	DENV-3: FFDLPLPWT (211-219)				
	11	DENV-4: FFDLPLPWL (213-221)				
	12	DENV-1,-2,-3: ILGDTAWDF (414-422/414-422/412-420)	$CD8^+$	B*07	DENV-2, -4	Simmons *et al.*, 2005
	13	DENV-4: ILGETAWDF (414-422)				

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
NS3	1	DENV-3: SVKKDLISY (71-79)	$CD8^+$	B*62	DENV-3	Zivny *et al.*, 1999
	2	DENV-3: IRYQTTATK (241-249)	$CD4^+$ CTL	DRB*1501[£]	DENV-3	Zeng *et al.*, 1996
	3	DENV-3: RKYLPAIVRE (202-211)				
	4	DENV-2: GLRTLIAPTRVVAA (215-229)	Not available	Not available	Not available	Simmons *et al.*, 2005
	5	DENV-2: IRYQTPAIRAEHTGR (240-254)				
	6	DENV-2: LSPVRVPNYNLIIMD (270-284)				
	7	DENV-2: VPNYNLIIMDEAHFT (275-289)				
	8	DENV-2: LIIMDEAHFTDPASI (280-294)				
	9	DENV-2: EAHFTDPASIAARGY (285-299)				
	10	DENV-2: EMGEAAGIFMTATPP (305-319)				
	11	DENV-2: AGIFMTATPPGSRDP (310-324)				
	12	DENV-2: KKVIQLSRKTFDSEY (380-394)				
	13	DENV-2: NDWDFVVTTDISEMG (400-414)				
	14	DENV-2: GDLPVWLAYRVAAEG (540-554)				
	15	DENV-2: KLKPRWLDARIYSDP (590-604)				
	16	DENV-2: WLDARIYSDPLALKE (595-609)				

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
NS3	17	DENV-1: GTSGSPIVNRE (133-143) DENV-1: GTSGSPIVNR (133-142)	$CD8^+$	A*11	Not available	Mongkolsapaya *et al.*, 2003
	18	DENV-2: AVSLDFSPGTSGSPI (125-139)	Not available	Not available		Simmons *et al.*, 2005
	19	DENV-2: GTSGSPIIDKK (133-143) DENV-2: GTSGSPIIDK (133-142)	$CD8^+$	A*11	Not available	Mongkolsapaya *et al.*, 2003
	20	DENV-2: GTSGSPIVDRK (133-143) DENV-2: GTSGSPIVDR (133-142)				
	21	DENV-2: GTSGSPIVDKK (133-143) DENV-2: GTSGSPIVDK (133-142)				
	22	DENV-3: GTSGSPIINRE (133-143)			DENV-3	Sanchez *et al.*, 2006
		DENV-3: GTSGSPIINRE (133-143) DENV-3: GTSGSPIINR (133-142)			Not available	Mongkolsapaya *et al.*, 2003
	23	DENV-4: GTSGSPIINRK (133-143) DENV-4: GTSGSPIINR (133-142)				
	24	DENV-1: NREGKIVGLYGNGVV (141-155)	$CD4^+$ (helper/cytotoxic not indicated)	DRB*1501	DENV-3	Mangada and Rothman, 2005
	25	DENV-2: DKKGKVVGLYGNGVV (141-155)				
	26	DENV-3: NREGKVVGLYGNGVV (141-155)				
	27	DENV-4: NRKGKIVGLYGNGVV (141-155)				
		DENV-4: VIGLYGNGV (146-154)	$CD4^+$ CTL		DENV-3	Kurane *et al.*, 1995
	28	DENV-1: RKLTIMDLHPGSGKT (186-200)	$CD4^+$ (helper/cytotoxic not indicated)	Not available	DENV-4	Mangada and Rothman, 2005
	29	DENV-2: RKLTIMDLHPGAGKT (186-200)				
	30	DENV-3: RNLTIMDLHPGSGKT (187-201)				
	31	DENV-4: KRLTIMDLHPGAGKT (186-200)				
	32	DENV-1: PTRVVASEMAEALKG (223-237)	$CD4^+$ (helper/cytotoxic not indicated)	DRB*1501	DENV-3	Mangada and Rothman, 2005
	33	DENV-3: PTRVVAAEMEEAMKG (224-238)				
		DENV-3: LAPTRVVAAEMEEAM (221-235)	$CD4^+$ CTL		DENV-3	Kurane *et al.*, 1998

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
NS3	34	DENV-2, -4: PTRVVAAEMEEALRG (223-237/223-237)	$CD4^+$ (helper/cytotoxic not indicated)	DRB*1501	DENV-3	Mangada and Rothman, 2005
		DENV-4: LAPTRVVAAEMEEAL (221-235) DENV-4: APTRVVAAE (222-230) DENV-4: LAPTRVVAAEME (221–232)	$CD8^+$	B*07	DENV-2	Mathew *et al.*, 1998,
		DENV-4: LAPTRVVAAEME (221-232)			DENV-4	Zivna *et al.*, 2002, Mathew *et al.*, 1998,
		DENV-4: LAPTRVVAAEMEEAL (221-235) DENV-4: VVAAEMEE (226-233) DENV-4: TRVVAAEMEEA (224-234)	$CD4^+$ CTL	DRB*1501	DENV-3	Kurane *et al.*, 1998
	35	DENV-2: LRGLPIRYQTPAIRA (235-249)	Not available	Not available	Not available	Simmons *et al.*, 2005
		DENV-2: EALRGLPIR (233-241)	$CD8^+$	A33	Not available	Simmons *et al.*, 2005, Loke *et al.*, 2001
	36	DENV-3: AMKGLPIRY (235-243)	$CD8^+$	B*62	DENV-3	Zivny *et al.*, 1999
	37	DENV-1: HTGKEIVDLMCHATF (252-266)	$CD4^+$ (helper/cytotoxic not indicated)	$DPw2^{\epsilon}$	DENV-3	Mangada and Rothman, 2005/ Kurane *et al.*, 1993
	38	DENV-2, -3, -4: HTGREIVDLMCHATF (251-265/252-266/251-265)				
		DENV-3: EIVDLMCHAT (255-264)	$CD4^+$ CTL	DPw2	DENV-3	Kurane *et al.*, 1993, Okamoto *et al.*, 1998
	39	DENV-1: WITDFPGKTVW (351-361)	$CD4^+$ CTL	DRB*1501	DENV-3	Zeng *et al.*, 1996
	40	DENV-2: TVWFVPSIK (358-366)	$CD8^+$	A*11	Not available	Simmons *et al.*, 2005, Loke *et al.*, 2001
	41	DENV-3: GNEWITDFVGKTVWF (348-362) DENV-3: WITDFVGKTVW (351-361)	$CD4^+$ CTL	DRB*1501	DENV-3	Zeng *et al.*, 1996

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
NS3	42	DENV-1, -3: TPEGIIPAL (500-508/500-508)	$CD8^+$	B35	DENV-4	Zivny *et al.*, 1995, Livingston *et al.*, 1995
	43	DENV-2: TPEGIIPSM (500-508)				
	44	DENV-4: TPEGIIPTL (500-508)				
	45	DENV-2: GESRKTFVE (527-535)	$CD8^+$	B*07	DENV-1	Zivna *et al.*, 2002
	46	DENV-3: GESRKTFVEL (528-537)		B60	DENV-3	Sanchez *et al.*, 2006
	47	DENV-1: FQYSDRRWCF (555-564)	$CD8^+$	A*24	DENV-2	Simmons *et al.*, 2005
	48	DENV-2: INYADRRWCF (555-564)			DENV-2, -4	Simmons *et al.*, 2005
		DENV-2: NYADRRWCF (556-564)			Not available	Loke *et al.*, 2001
	49	DENV-4: ISYKDREWCF (555-564)			DENV-4	Simmons *et al.*, 2005
	50	DENV-1: KEGERKKLRPRWLDA (584-598)	$CD4^+$ (helper/cytotoxic not indicated)	Not available	DENV-4	Mangada and Rothman, 2005
	51	DENV-2: EGERKKLKPRWLDAIY (585-599)	Not available		Not available	Simmons *et al.*, 2005
	52	DENV-2: KEGERKKLKPRWLDA (584-598)	$CD4^+$ (helper/cytotoxic not indicated)			Mangada and Rothman, 2005
	53	DENV-3: KEGEKKKLRPRWLDA (585-599)				
	54	DENV-4: REGEKKKLRPRW*DAR (584-599)				

Dengue protein	Epitope		Type of cellular immune response	HLA restricting molecule	Serotype of infection or of vaccine received	Reference (s)
	#	Sequence (amino acid position)				
NS4a	1	DENV-1: MLLALIAVL (56-64)	$CD8^+$	A*0201	DENV-1, -2, -3	Bashyam *et al.*, 2006
	2	DENV-2: LLLTLLATV (56-64)				
	3	DENV-3: LLLGLMILL (56-64)				
	4	DENV-4: MLVALLGAM (56-64)				
	5	DENV-2: LATVTGGIFLFLMSGRGIGK (61-80)	Not available	Not available	Not available	Simmons *et al.*, 2005
NS4b	1	DENV-1: VLMLVAHYA (112-120)	$CD8^+$	A*0201	DENV-1, -2, -3	Bashyam *et al.*, 2006
	2	DENV-2: FLLVAHYAI (112-120)				
	3	DENV-3: VLLLVTHYA (111-119)				
	4	DENV-4: LVMLLVHYA (108-116)				
	5	DENV-1: ILLMRTTWA (182-190)				
	6	DENV-2: VLLMRTTWA (181-189)				
	7	DENV-3: LLLMRTSWA (181-189)				
	8	DENV-4: LLLMRTTWA (178-186)				
NS5	1	DENV-2: DVFFTPPEK (131-139)	$CD8^+$	A*11	Not available	Loke *et al.*, 2001
	2	DENV-2: YILRDVSKK (517-525)				

¶: This refers to the type of T-cell (either $CD4^+$ or $CD8^+$) used in the experiment to study the immunogenicity of the DENV peptide.
¥: This refers to the serotype of the DENV vaccine received by the human subject or the serotype of the virus that infected the patient.
€: DPB1*0201-02 is serologically defined as DPw2
£: The HLA allele DRB1*1501 is serologically defined as DR15 or DR2
¢: DPB1*0401-02 is serologically defined as DPw4
*: The undergraduate attachment student to our lab, Mr. Lam Jian Hang, assisted with the construction of the catalogue and analysis of the data therein.

Appendix 2: Annotation errors in DV records collected from the NCBI Entrez protein database.

DV1 entry	Error/discrepancy description
AAK29447	• The position of the C terminal end of the protein NS4B is mis-annotated. Instead of 2293 it should be 2493. Evidence by sequence similarity to other strains. • The position of the N-terminal of the protein NS5 is mis-annotated. Instead of 2294 it should be 2494. Evidence by sequence similarity to other strains.
A42551	• The C-terminal position of the precursor membrane protein region has been mis-annotated. It should be 205..280 instead of 205..281. Evidence by sequence similarity to other strains. • The N-terminal position of the envelope protein region has been mis-annotated. It should be 281..774 instead of 282..774. Evidence by sequence similarity to other strains. • The C-terminal position of the NS1 protein region has been mis-annotated. It should be 775..1126 instead of 775..1127. Evidence by sequence similarity to other strains. • The N-terminal position of the NS2a protein region has been mis-annotated. It should be 1127..1344 instead of 1128..1344. Evidence by sequence similarity to other strains.
P33478	• The C-terminal position of the NS1 protein region has been mis-annotated. It should be 775..1126 instead of 775..1127. Evidence for this can be found in Fu *et al*., (1992). • The N-terminal position of the NS2a protein region has been mis-annotated. It should be 1127..1344 instead of 1128..1344. Evidence for this can be found in Fu *et al*., (1992).
AAN03445	• Under the field "Features", we see the following statement: Protein 1..3392 /product="envelope glycoprotein" The statement is not correct because envelope glycoprotein is only a part of the whole polyprotein. Instead of envelope glycoprotein it should have been "polyprotein" as in entry AAO47361 (GI:34596500)
AAB70694	• The N-terminal position of the capsid protein region has been mis-annotated. It should be 1..114 instead of 2..114. Evidence by sequence similarity to other strains.
AAB70696	• The N-terminal position of the capsid protein region has been mis-annotated. It should be 1..114 instead of 2..114. Evidence by sequence similarity to other strains. • The C-terminal position of the precursor membrane protein region has been mis-annotated. It should be 115..280 instead of 115..278. Evidence by sequence similarity to other strains. • The C-terminal position of the mature membrane protein region has been mis-annotated. It should be 206..280 instead of 206..278. Evidence by sequence similarity to other strains. • The N-terminal position of the envelope protein region has been mis-annotated. It should be 281..775 instead of 279..775.

	Evidence by sequence similarity to other strains.
AAB70695	• The N-terminal position of the capsid protein region has been mis-annotated. It should be 1..114 instead of 2..114. Evidence by sequence similarity to other strains. • The C-terminal position of the precursor membrane protein region has been mis-annotated. It should be 115..280 instead of 115..278. Evidence by sequence similarity to other strains. • The C-terminal position of the mature membrane protein region has been mis-annotated. It should be 206..280 instead of 206..278. Evidence by sequence similarity to other strains. • The N-terminal position of envelope protein region has been mis-annotated. It should be 281..775 instead of 279..775. Evidence by sequence similarity to other strains.
DV2 entry	**Error/discrepancy description**
AAL00888	• The position of the C-terminal of the capsid protein is mis-annotated. It should be 1..114 instead of 1..150. Evidence by sequence similarity to other strains. • The position of the N-terminal of the precursor membrane protein is mis-annotated. It should be 115..280 instead of 151..280. Evidence by sequence similarity to other strains.
CAD31751	• The protein 115..280 should be annotated as prM protein instead of envelope protein. • The protein 281..774 should be annotated as envelope protein instead of prM protein. • The position of the C-terminal of the envelope protein is mis-annotated. It should be 281..775 instead of 281..774. Evidence by sequence similarity to other strains. • The position of the N-terminal of the NS1 protein is mis-annotated. It should be 776..1127 instead of 775..1127. Evidence by sequence similarity to other strains. • The position of the C-terminal of the envelope protein is mis-annotated. It should be 1346..1475 instead of 1346..1474. Evidence by sequence similarity to other strains. • The position of the N-terminal of the NS3 protein is mis-annotated. It should be 1476..2093 instead of 1475..2093. Evidence by sequence similarity to other strains.
AAA42941	• The N-terminal position of the capsid protein region has been mis-annotated. It should be 1..114 instead of 2..114. Evidence by sequence similarity to other strains. • The position of the C-terminal of the NS4a protein is mis-annotated. It should be 2094..2243 instead of 2094..2379. Evidence by sequence similarity to other strains. • The position of the N-terminal of the NS4b protein is mis-annotated. It should be 2244..2491 instead of 2380..2491. Evidence by sequence similarity to other strains.
DV3 entry	**Error/discrepancy description**
P27915, AAA99437 and GNWVD3	• All the three entries have NS1/NS2a and NS4a/NS4b junctions mis-annotated; the amino acid sequence at these junctions in the three entries were not similar to those

	described in Osatomi *et al.* (1990).
DV4 entry	**Error/discrepancy description**
AAB28474	• The position of the C-terminal of the capsid protein is mis-annotated. It should be 1..113 instead of 1..110. Evidence by sequence similarity to other strains. • The position of the N-terminal of the precursor membrane protein is mis-annotated. It should be 114..279 instead of 112..279. Evidence by sequence similarity to other strains.
AAA42964	• The position of the C-terminal of the envelope protein is mis-annotated. It should be 280..774 instead of 280..733. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS1 protein is mis-annotated. It should be 775..1126 instead of 734..1184. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS2a protein is mis-annotated. It should be 1127..1344 instead of 1185..1343. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS2b protein is mis-annotated. It should be 1345..1474 instead of 1344..1473. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS3 protein is mis-annotated. It should be 1475..2092 instead of 1474..2091. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS4a protein is mis-annotated. It should be 2093..2242 instead of 2092..2374. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS4b protein is mis-annotated. It should be 2243..2487 instead of 2092..2374. Evidence by sequence similarity to other strains. • The position of both the N and C termini of the NS5 protein is mis-annotated. It should be 2488..3387 instead of 2487..3386. Evidence by sequence similarity to other strains.
P09866	• The position of the C-terminus of the NS1 protein is mis-annotated. It should be 775..1126 instead of 775..1185. Evidence by sequence similarity to other strains. • The position of the N-terminus of the NS2a protein is mis-annotated. It should be 1127..1344 instead of 1186..1344. Evidence by sequence similarity to other strains. • The position of the C-terminus of the NS4a protein is mis-annotated. It should be 2093..2242 instead of 2093..2375. Evidence by sequence similarity to other strains. • The position of the N-terminus of the NS4b protein is mis-annotated. It should be 2243..2487 instead of 2376..2487. Evidence by sequence similarity to other strains.
GNWVDF	• The position of the C-terminus of the NS1 protein is mis-annotated. It should be 774..1125 instead of 774..1184. Evidence by sequence similarity to other strains. • The position of the N-terminus of the NS2a protein is mis-annotated. It should be 1126..1343 instead of 1185..1343.

	Evidence by sequence similarity to other strains. • The position of the C-terminus of the NS4a protein is mis-annotated. It should be 2092..2241 instead of 2092..2374. Evidence by sequence similarity to other strains. • The position of the N-terminus of the NS4b protein is mis-annotated. It should be 2242..2486 instead of 2375..2486. Evidence by sequence similarity to other strains.

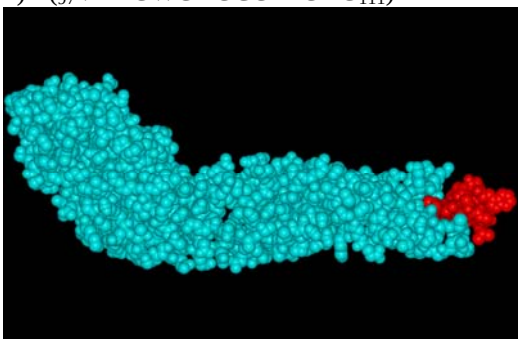
References

Osatomi K, Sumiyoshi H: **Complete nucleotide sequence of dengue type 3 virus genome RNA**. *Virology* 1990, **176**(2):643-647.

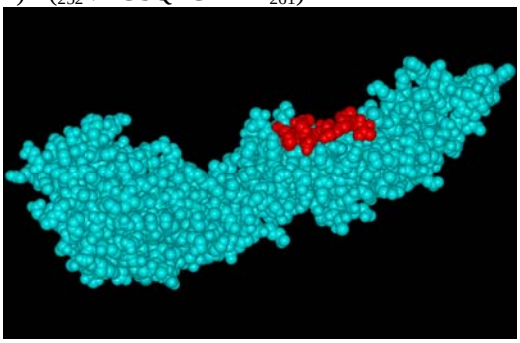
Fu J, Tan BH, Yap EH, Chan YC, Tan YH: **Full-length cDNA sequence of dengue type 1 virus (Singapore strain S275/90)**. *Virology* 1992, **188**(2):953-958.

Appendix 3: Molecular location of 19 pan-DENV sequences (in red) on the protein's 3-D structure. These sequences were mapped on the available crystallographic models of the E ectodomain (PDB Accession No. 1OAN; 394 out of 493-495 residues), NS3 (1BEF and 2BMF, 181 and 451 out of 618-619 residues, respectively) and NS5 fragments (1R6A, 295 out of 900-904 residues). The major portions of eleven of the 19 pan-DENV sequences were buried (NS3-$_{148}$GLYGNGVVT$_{156}$, $_{256}$EIVDLMCHATFT$_{267}$, $_{284}$MDEAHFTDP$_{292}$, $_{296}$AARGYISTRV$_{305}$, $_{313}$IFMTATPPG$_{321}$, $_{357}$GKTVWFVPSIK$_{367}$, $_{406}$VVTTDISEMGANF$_{418}$, and $_{491}$EAKMLLDNI$_{499}$; NS5-$_{79}$DLGCGRGGWSYY$_{90}$, $_{141}$DTLLCDIGESS$_{151}$ and $_{209}$PLSRNSTHEMYW$_{220}$), 2 were partially buried/exposed (NS3-$_{46}$FHTMWHVTRG$_{55}$ and $_{537}$LMRRGDLPVWL$_{547}$) and the remaining 6 were exposed (E-$_{97}$VDRGWGNGCGLFGKG$_{111}$ and $_{252}$VLGSQEGAMH$_{261}$; NS3-$_{189}$LTIMDLHPG$_{197}$ and $_{383}$VIQLSRKTFD$_{392}$; NS5-$_{6}$GETLGEKWK$_{14}$ and $_{104}$TKGGPGHEEP$_{113}$) at the surface of the corresponding structures.

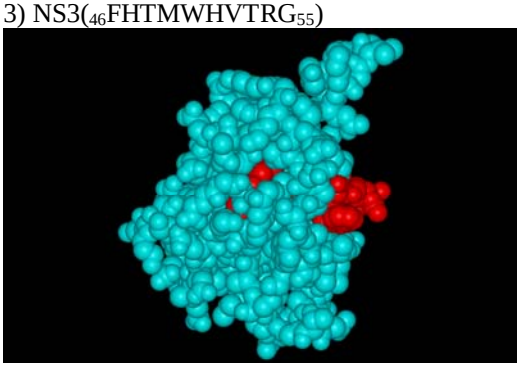
1) E($_{97}$VDRGWGNGCGLFGKG$_{111}$)



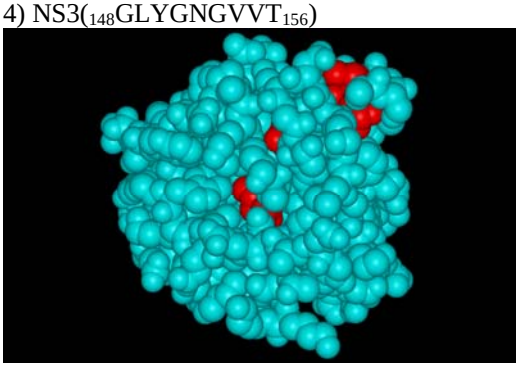
2) E($_{252}$VLGSQEGAMH$_{261}$)



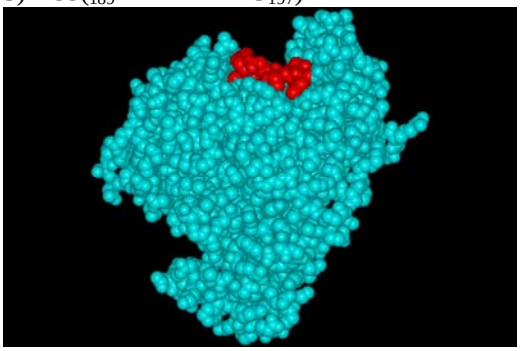
3) NS3($_{46}$FHTMWHVTRG$_{55}$)



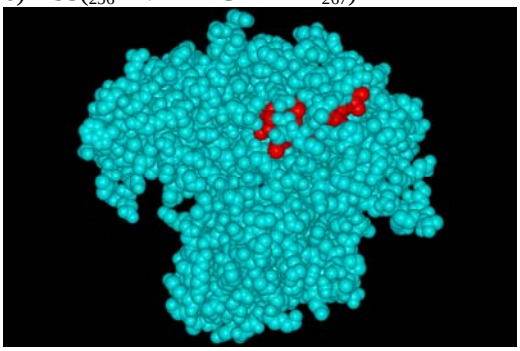
4) NS3($_{148}$GLYGNGVVT$_{156}$)



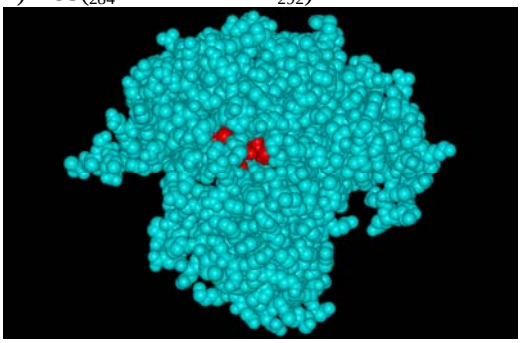
5) NS3($_{189}$LTIMDLHPG$_{197}$)



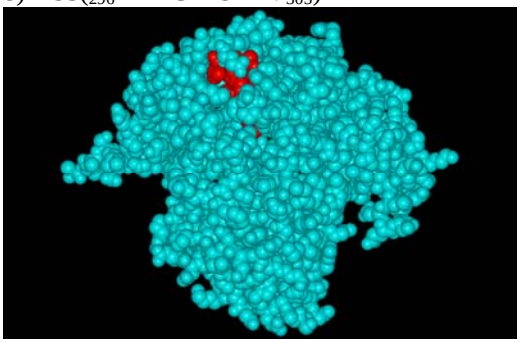
6) NS3($_{256}$EIVDLMCHATFT$_{267}$)



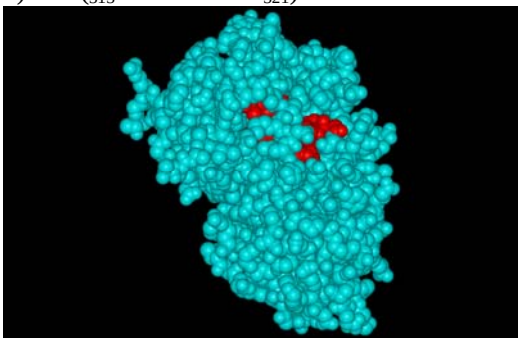
7) NS3($_{284}$MDEAHFTDP$_{292}$)



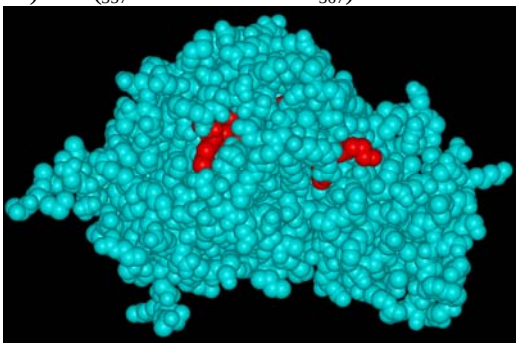
8) NS3($_{296}$AARGYISTRV$_{305}$)



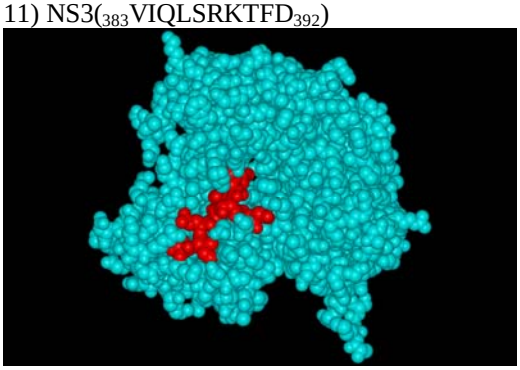
9) NS3($_{313}$IFMTATPPG$_{321}$)



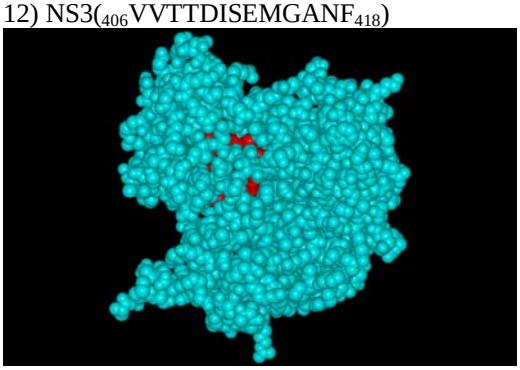
10) NS3($_{357}$GKTVWFVPSIK$_{367}$)



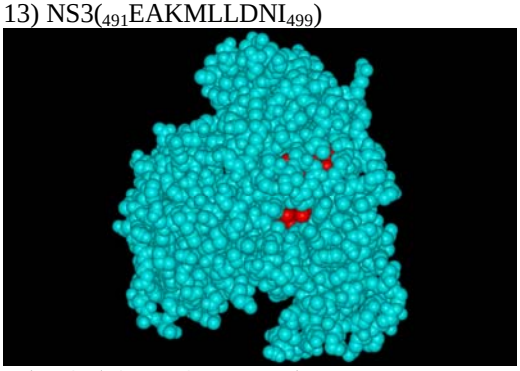
11) NS3($_{383}$VIQLSRKTFD$_{392}$)



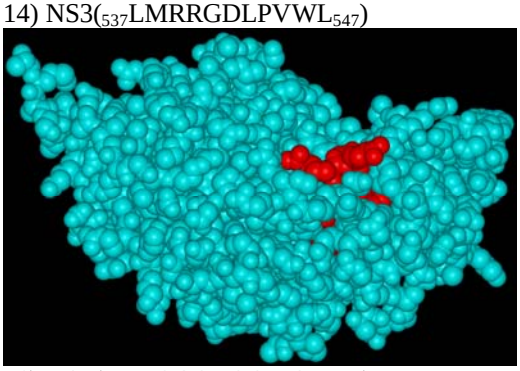
12) NS3($_{406}$VVTTDISEMGANF$_{418}$)



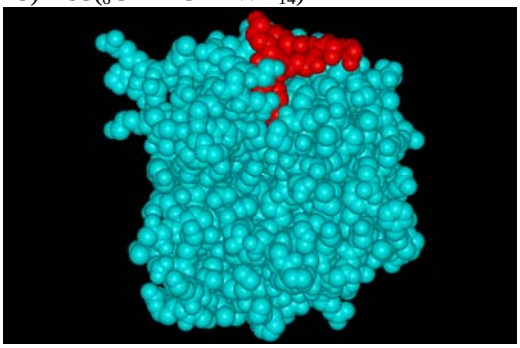
13) NS3($_{491}$EAKMLLDNI$_{499}$)



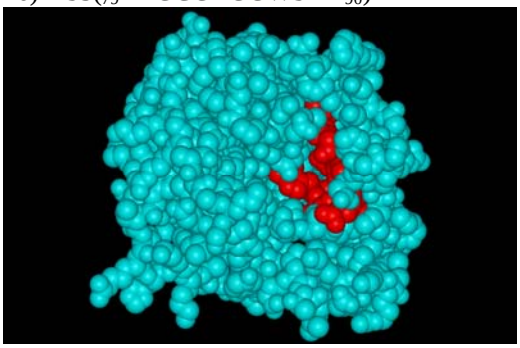
14) NS3($_{537}$LMRRGDLPVWL$_{547}$)



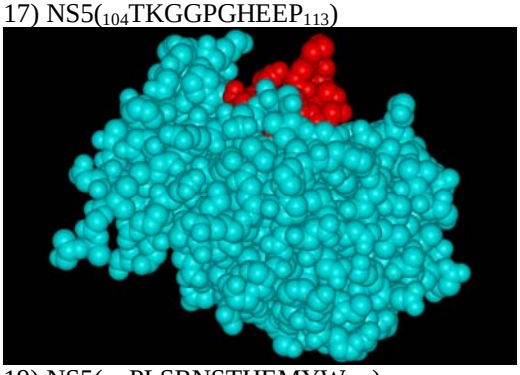
15) NS5($_{6}$GETLGEKWK$_{14}$)



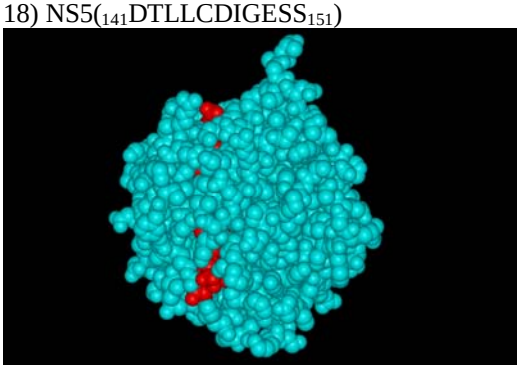
16) NS5($_{79}$DLGCGRGGWSYY$_{90}$)



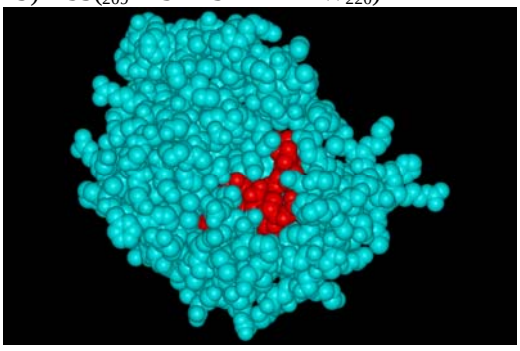
17) NS5($_{104}$TKGGPGHEEP$_{113}$)



18) NS5($_{141}$DTLLCDIGESS$_{151}$)



19) NS5($_{209}$PLSRNSTHEMYW$_{220}$)



Appendix 4: Candidate putative HLA supertype-restricted binding nonamer peptides in pan-DENV sequences, predicted by immunoinformatic algorithms.

DENV protein	Pan-DENV sequence and the predicted HLA supertype-restricted binding nonamer(s) [a]	HLA supertype-restriction of predicted nonamer peptide[b]				
		Class I[c]			Class II[d]	
		NetCTL	Multipred	ARB	Multipred	TEPITOPE
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$					
	$_{99}$RGWGNGCGL$_{107}$	B7	-	-	-	-
	$_{100}$GWGNGCGLF$_{108}$	A24	-	-	-	-
NS1	$_{12}$ELKCGSGIF$_{20}$					
	$_{12}$ELKCGSGIF$_{20}$	A26, B8, B62	-	-	-	-
	$_{25}$VHTWTEQYKFQ$_{35}$					
	$_{26}$HTWTEQYKF$_{34}$	A1, A24, A26, B8, B27, B58, B62	-	-	-	-
	$_{193}$AVHADMGYWIES$_{204}$					
	$_{193}$AVHADMGYW$_{201}$	A26, B58	-	-	-	-
	$_{194}$VHADMGYWI$_{202}$	A24, B39	-	-	DR	-
	$_{195}$HADMGYWIE$_{203}$	A1	-	-	-	-
	$_{229}$HTLWSNGVLES$_{239}$					
	$_{229}$HTLWSNGVL$_{237}$	A1, B8, B39, B62	-	-	-	-
	$_{231}$LWSNGVLES$_{239}$	-	-	-	-	DR
	$_{325}$GEDGCWYGMEIRP$_{337}$					
	$_{325}$GEDGCWYGM$_{333}$	B44	-	-	-	-
	$_{328}$GCWYGMEIR$_{336}$	-	-	A3	-	-
NS3	$_{46}$FHTMWHVTRG$_{55}$					
	$_{46}$FHTMWHVTR$_{54}$	B39	A3	-	DR	-
	$_{47}$HTMWHVTRG$_{55}$	-	A3	-	-	-
	$_{189}$LTIMDLHPG$_{197}$					
	$_{189}$LTIMDLHPG$_{197}$	-	-	-	**DR**	**DR**
	$_{256}$EIVDLMCHATFT$_{267}$					
	$_{256}$EIVDLMCHA$_{264}$	A26	A2	-	-	-
	$_{257}$IVDLMCHAT$_{265}$	-	-	-	DR	-

	$_{258}$VDLMCHATF$_{266}$	B8, **B44**	-	**B44**	DR	-
	$_{259}$DLMCHATFT$_{267}$	-	**A2**	**A2**	-	-
	$_{296}$AARGYISTRV$_{305}$					
	$_{296}$AARGYISTR$_{304}$	**A3**	**A3**	**A3**	-	-
	$_{297}$ARGYISTRV$_{305}$	B27	-	-	-	-
	$_{313}$IFMTATPPG$_{321}$					
	$_{313}$IFMTATPPG$_{321}$	-	-	-	**DR**	**DR**
	$_{357}$GKTVWFVPSIK$_{367}$					
	$_{358}$KTVWFVPSI$_{366}$	**A2**, A24, A26, B58	**A2**	**A2**	-	-
	$_{359}$TVWFVPSIK$_{367}$	**A3**	**A3**	**A3**	-	-
	$_{383}$VIQLSRKTFD$_{392}$					
	$_{383}$VIQLSRKTF$_{391}$	B7, B8, B62	-	-	DR	-
	$_{384}$IQLSRKTFD$_{392}$	-	-	-	DR	-
	$_{406}$VVTTDISEMGANF$_{418}$					
	$_{406}$VVTTDISEM$_{414}$	A26, B62	-	-	DR	-
	$_{407}$VTTDISEMG$_{415}$	-	-	-	DR	-
	$_{408}$TTDISEMGA$_{416}$	A1	-	-	-	-
	$_{410}$DISEMGANF$_{418}$	A1, A26, B62	-	-	-	-
	$_{537}$LMRRGDLPVWL$_{547}$					
	$_{537}$LMRRGDLPV$_{545}$	**A2**, B8, B62	-	**A2**	DR	-
	$_{538}$MRRGDLPVW$_{546}$	B27	-	-	-	-
	$_{539}$RRGDLPVWL$_{547}$	B27, B39	-	-	-	-
NS4a	$_{126}$QRTPQDNQL$_{134}$					
	$_{126}$QRTPQDNQL$_{134}$	B27, B39	-	-	-	-
NS4b	$_{35}$PASAWTLYAVATT$_{47}$					
	$_{36}$ASAWTLYAV$_{44}$	A1, **A2**	**A2**	**A2**	-	-
	$_{37}$SAWTLYAVA$_{45}$	-	**A2**	**A2**	-	-
	$_{39}$WTLYAVATT$_{47}$	-	**A2**	**A2**	DR	-
	$_{118}$HYAIIGPGLQAKATREAQKR$_{137}$					
	$_{118}$HYAIIGPGL$_{126}$	A24, B39	-	-	-	-
	$_{119}$YAIIGPGLQ$_{127}$	-	-	-	**DR**	**DR**
	$_{120}$AIIGPGLQA$_{128}$	-	A3	-	-	-
	$_{121}$IIGPGLQAK$_{129}$	**A3**	**A3**	-	-	-
	$_{126}$LQAKATREA$_{134}$	B62	-	A2	DR	-

	$_{127}$QAKATREAQ$_{135}$	B8	-	-	-	-
	$_{128}$AKATREAQK$_{136}$	B27	A3	-	-	-
	$_{129}$KATREAQKR$_{137}$	-	A3	-	-	-
	$_{139}$AAGIMKNPTVDGI$_{151}$					
	$_{142}$IMKNPTVDG$_{150}$	-	A3	-	DR	-
	$_{143}$MKNPTVDGI$_{151}$	-	-	B7	DR	-
	$_{223}$ANIFRGSYLAGAGL$_{236}$					
	$_{223}$ANIFRGSYL$_{231}$	B7	-	A2	-	-
	$_{224}$NIFRGSYLA$_{232}$	**A2**	**A3**	**A2, A3**	-	-
	$_{225}$IFRGSYLAG$_{233}$	-	-	-	DR	-
	$_{226}$FRGSYLAGA$_{234}$	B27	**A2**	**A2**	**DR**	**DR**
	$_{228}$GSYLAGAGL$_{236}$	B39, B44, B62	-	-	-	-
NS5	$_{6}$GETLGEKWK$_{14}$					
	$_{6}$GETLGEKWK$_{14}$	B44	-	-	-	-
	$_{79}$DLGCGRGGWSYY$_{90}$					
	$_{81}$GCGRGGWSY$_{89}$	A1, B62	-	-	-	-
	$_{82}$CGRGGWSYY$_{90}$	A1, A26, B62	-	-	-	-
	$_{141}$DTLLCDIGESS$_{151}$					
	$_{142}$TLLCDIGES$_{150}$	-	-	A2	-	-
	$_{143}$LLCDIGESS$_{151}$	-	-	-	DR	-
	$_{209}$PLSRNSTHEMYW$_{220}$					
	$_{210}$LSRNSTHEM$_{218}$	**B7**, B58, B62	-	**B7**	DR	-
	$_{211}$SRNSTHEMY$_{219}$	A1, B8, B27	-	-	-	-
	$_{212}$RNSTHEMYW$_{220}$	B58	-	-	-	-
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$					
	$_{343}$MTDTTPFGQ$_{351}$	A1	-	-	-	-
	$_{345}$DTTPFGQQR$_{353}$	-	**A3**	**A3**	-	-
	$_{346}$TTPFGQQRV$_{354}$	A1, A26	-	-	-	-
	$_{347}$TPFGQQRVF$_{355}$	**B7**, B8	-	**B7**	-	-
	$_{348}$PFGQQRVFK$_{356}$	-	A3	-	-	-
	$_{349}$FGQQRVFKE$_{357}$	-	-	-	DR	-
	$_{350}$GQQRVFKEK$_{358}$	**A3**, B27	**A3**	-	-	-
	$_{354}$VFKEKVDTR$_{362}$	-	A3	-	-	-
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$					
	$_{450}$CVYNMMGKR$_{458}$	**A3**	**A3**	**A3**	-	-

$_{451}$VYNMMGKRE$_{459}$	-	-	-	DR	-
$_{452}$YNMMGKREK$_{460}$	-	**A3**	**A3**	**DR**	**DR**
$_{453}$NMMGKREKK$_{461}$	**A3**	**A3**	**A3**	-	-
$_{454}$MMGKREKKL$_{462}$	B8	A2	-	-	-
$_{457}$KREKKLGEF$_{465}$	A1, B8, B27	-	-	-	-
$_{458}$REKKLGEFG$_{466}$	-	-	B44	-	-
$_{468}$AKGSRAIWYMWLGAR$_{482}$					
$_{469}$KGSRAIWYM$_{477}$	B58	-	-	-	-
$_{470}$GSRAIWYMW$_{478}$	B58	-	-	-	-
$_{471}$SRAIWYMWL$_{479}$	B27, B39	-	-	-	-
$_{473}$AIWYMWLGA$_{481}$	**A2**	-	**A2**	-	-
$_{474}$IWYMWLGAR$_{482}$	-	-	-	DR	-
$_{531}$YADDTAGWDTRIT$_{543}$					
$_{531}$YADDTAGWD$_{539}$	-	-	-	DR	-
$_{534}$DTAGWDTRI$_{542}$	A1, A26	A2	-	-	-
$_{568}$IFKLTYQNKVV$_{578}$					
$_{568}$IFKLTYQNK$_{576}$	A3, A24	-	-	-	-
$_{569}$FKLTYQNKV$_{577}$	**A2**	**A2**	**A2**	DR	-
$_{570}$KLTYQNKVV$_{578}$	**A2**	**A2**	-	-	-
$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$					
$_{599}$RGSGQVGTY$_{607}$	A1, B58, B62	-	-	-	-
$_{601}$SGQVGTYGL$_{609}$	B39	-	-	-	-
$_{604}$VGTYGLNTF$_{612}$	B58, B62	-	-	DR	-
$_{605}$GTYGLNTFT$_{613}$	-	-	A2	-	-
$_{606}$TYGLNTFTN$_{614}$	A24	-	-	-	-
$_{607}$YGLNTFTNM$_{615}$	A26	-	-	DR	-
$_{658}$RMAISGDDCVVKP$_{670}$					
$_{659}$MAISGDDCV$_{667}$	-	-	A2, B7	-	-
$_{660}$AISGDDCVV$_{668}$	**A2**	**A2**	**A2**	-	-
$_{661}$ISGDDCVVK$_{669}$	**A3**	**A3**	-	-	-
$_{707}$VPFCSHHFH$_{715}$					
$_{707}$VPFCSHHFH$_{715}$	-	-	A3	DR	-
$_{765}$LMYFHRRDLRLA$_{776}$					
$_{765}$LMYFHRRDL$_{773}$	B39, B8, B62	-	-	**DR**	**DR**
$_{766}$MYFHRRDLR$_{774}$	**A3**	**A3**	**A3**	DR	-
$_{767}$YFHRRDLRL$_{775}$	A1, A24, B8, B39	A2	-	DR	-

$_{768}$FHRRDLRLA$_{776}$	-	-	-	DR	-
$_{790}$PTSRTTWSIHA$_{800}$					
$_{790}$PTSRTTWSI$_{798}$	A1, A24	A2	-	-	-
$_{792}$SRTTWSIHA$_{800}$	B27	-	-	-	-

[a] Amino acid positions of the pan-DENV sequences and the predicted nonamers are numbered according to the sequence alignments of the 4 DENV types

[b] HLA supertype-restrictions that were predicted by at least two prediction models are highlighted in bold

[c] Peptides specific to HLA class I supertypes were predicted by use of NetCTL (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 and B62), ARB (A2, A3, B44, and B7) and Multipred (A2 and A3)

[d] Sequences identified as specific to class II were predicted by use of TEPITOPE (DR) and Multipred (DR) as described in methods

Appendix 5: Intra-type representation of candidate putative HLA supertype-restricted nonamer peptides predicted by immunoinformatics algorithms.

DENV protein	Pan-DENV sequence and the putative HLA supertype-restricted nonamer peptide(s)[a]	Intra-type representation (%)[b] and total sequences analyzed (#)[c]			
		DENV-1	DENV-2	DENV-3	DENV-4
E	$_{97}$VDRGWGNGCGLFGKG$_{111}$				
	$_{99}$RGWGNGCGL$_{107}$	98%(580)	98%(811)	100%(372)	100%(320)
	$_{100}$GWGNGCGLF$_{108}$	99%(580)	98%(811)	100%(372)	95%(320)
NS1	$_{12}$ELKCGSGIF$_{20}$				
	$_{12}$ELKCGSGIF$_{20}$	99%(366)	95%(603)	83%(201)	99%(141)
	$_{25}$VHTWTEQYKFQ$_{35}$				
	$_{26}$HTWTEQYKF$_{34}$	99%(350)	96%(555)	100%(197)	95%(140)
	$_{193}$AVHADMGYWIES$_{204}$				
	$_{193}$AVHADMGYW$_{201}$	100%(104)	95%(197)	98%(117)	100%(28)
	$_{194}$VHADMGYWI$_{202}$	100%(104)	96%(197)	98%(117)	96%(28)
	$_{195}$HADMGYWIE$_{203}$	100%(104)	97%(197)	98%(117)	96%(28)
	$_{229}$HTLWSNGVLES$_{239}$				
	$_{229}$HTLWSNGVL$_{237}$	98%(124)	97%(215)	97%(117)	100%(28)
	$_{231}$LWSNGVLES$_{239}$	97%(126)	97%(215)	98%(117)	100%(28)
	$_{325}$GEDGCWYGMEIRP$_{337}$				
	$_{325}$GEDGCWYGM$_{333}$	98%(112)	98%(202)	98%(113)	100%(28)
	$_{328}$GCWYGMEIR$_{336}$	100% (104)	99%(197)	100%(115)	100%(28)
NS3	$_{46}$FHTMWHVTRG$_{55}$				
	$_{46}$FHTMWHVTR$_{54}$	100%(89)	100%(132)	100%(68)	100%(28)
	$_{47}$HTMWHVTRG$_{55}$	100%(89)	100%(132)	100%(68)	100%(28)
	$_{189}$LTIMDLHPG$_{197}$				
	$_{189}$LTIMDLHPG$_{197}$	100%(97)	99%(141)	100%(97)	100%(29)
	$_{256}$EIVDLMCHATFT$_{267}$				
	$_{256}$EIVDLMCHA$_{264}$	99%(98)	100%(137)	100%(103)	100%(29)
	$_{257}$IVDLMCHAT$_{265}$	99%(98)	100%(137)	100%(103)	100%(29)
	$_{258}$VDLMCHATF$_{266}$	99%(98)	100%(137)	100%(103)	100%(29)
	$_{259}$DLMCHATFT$_{267}$	99%(98)	100%(137)	100%(103)	100%(29)
	$_{296}$AARGYISTRV$_{305}$				
	$_{296}$AARGYISTR$_{304}$	97%(90)	100%(135)	97%(74)	100%(27)
	$_{297}$ARGYISTRV$_{305}$	97%(90)	100%(135)	97%(74)	100%(27)
	$_{313}$IFMTATPPG$_{321}$				
	$_{313}$IFMTATPPG$_{321}$	100%(90)	100%(135)	100%(74)	100%(27)
	$_{357}$GKTVWFVPSIK$_{367}$				
	$_{358}$KTVWFVPSI$_{366}$	99%(90)	100%(135)	99%(181)	96%(27)
	$_{359}$TVWFVPSIK$_{367}$	99%(90)	100%(135)	100%(181)	100%(27)
	$_{383}$VIQLSRKTFD$_{392}$				
	$_{383}$VIQLSRKTF$_{391}$	81%(90)	99%(135)	99%(181)	100%(27)
	$_{384}$IQLSRKTFD$_{392}$	81%(90)	99%(135)	98%(181)	100%(27)
	$_{406}$VVTTDISEMGANF$_{418}$				
	$_{406}$VVTTDISEM$_{414}$	98%(90)	99%(135)	98%(181)	100%(27)
	$_{407}$VTTDISEMG$_{415}$	98%(90)	99%(135)	99%(181)	100%(27)
	$_{408}$TTDISEMGA$_{416}$	98%(90)	100%(135)	99%(181)	100%(27)
	$_{410}$DISEMGANF$_{418}$	98%(90)	100%(135)	99%(181)	100%(27)
	$_{537}$LMRRGDLPVWL$_{547}$				
	$_{537}$LMRRGDLPV$_{545}$	99%(89)	100%(133)	99%(181)	93%(27)
	$_{538}$MRRGDLPVW$_{546}$	99%(89)	100%(133)	99%(181)	93%(27)
	$_{539}$RRGDLPVWL$_{547}$	100%(89)	100%(133)	99%(181)	93%(27)
NS4a	$_{126}$QRTPQDNQL$_{134}$				
	$_{126}$QRTPQDNQL$_{134}$	98%(87)	100%(126)	100%(70)	100%(26)

NS4b	$_{35}$PASAWTLYAVATT$_{47}$				
	$_{36}$ASAWTLYAV$_{44}$	100%(89)	100%(127)	100%(70)	100%(27)
	$_{37}$SAWTLYAVA$_{45}$	100%(89)	100%(127)	100%(70)	100%(27)
	$_{39}$WTLYAVATT$_{47}$	100%(89)	100%(127)	100%(70)	100%(27)
	$_{118}$HYAIIGPGLQAKATREAQKR$_{137}$				
	$_{118}$HYAIIGPGL$_{126}$	100%(89)	97%(127)	100%(70)	98%(109)
	$_{119}$YAIIGPGLQ$_{127}$	100%(89)	97%(127)	100%(70)	98%(109)
	$_{120}$AIIGPGLQA$_{128}$	100%(89)	97%(127)	100%(70)	98%(109)
	$_{121}$IIGPGLQAK$_{129}$	100%(89)	97%(127)	100%(70)	99%(109)
	$_{126}$LQAKATREA$_{134}$	99%(89)	97%(127)	100%(70)	100%(109)
	$_{127}$QAKATREAQ$_{135}$	99%(89)	95%(127)	100%(70)	100%(109)
	$_{128}$AKATREAQK$_{136}$	99%(89)	95%(127)	100%(70)	100%(109)
	$_{129}$KATREAQKR$_{137}$	99%(89)	95%(127)	100%(70)	100%(109)
	$_{139}$AAGIMKNPTVDGI$_{151}$				
	$_{142}$IMKNPTVDG$_{150}$	96%(89)	98%(127)	97%(70)	100%(109)
	$_{143}$MKNPTVDGI$_{151}$	98%(89)	98%(127)	97%(70)	100%(109)
	$_{223}$ANIFRGSYLAGAGL$_{236}$				
	$_{223}$ANIFRGSYL$_{231}$	100%(87)	100%(129)	100%(70)	98%(109)
	$_{224}$NIFRGSYLA$_{232}$	100%(87)	100%(129)	100%(70)	100%(109)
	$_{225}$IFRGSYLAG$_{233}$	100%(87)	100%(129)	100%(70)	100%(109)
	$_{226}$FRGSYLAGA$_{234}$	100%(87)	100%(129)	97%(70)	100%(109)
	$_{228}$GSYLAGAGL$_{236}$	100%(87)	100%(129)	97%(70)	100%(109)
NS5	$_{6}$GETLGEKWK$_{14}$				
	$_{6}$GETLGEKWK$_{14}$	92%(87)	98%(131)	100%(70)	100%(109)
	$_{79}$DLGCGRGGWSYY$_{90}$				
	$_{81}$GCGRGGWSY$_{89}$	100%(87)	98%(130)	100%(78)	100%(27)
	$_{82}$CGRGGWSYY$_{90}$	100%(87)	98%(130)	100%(80)	100%(27)
	$_{141}$DTLLCDIGESS$_{151}$				
	$_{142}$TLLCDIGES$_{150}$	100%(87)	100%(130)	100%(74)	100%(27)
	$_{143}$LLCDIGESS$_{151}$	100%(87)	100%(130)	100%(74)	100%(27)
	$_{209}$PLSRNSTHEMYW$_{220}$				
	$_{210}$LSRNSTHEM$_{218}$	100%(87)	100%(130)	100%(70)	100%(27)
	$_{211}$SRNSTHEMY$_{219}$	100%(87)	100%(130)	100%(70)	100%(27)
	$_{212}$RNSTHEMYW$_{220}$	100%(87)	100%(130)	100%(70)	100%(27)
	$_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$				
	$_{343}$MTDTTPFGQ$_{351}$	100%(87)	98%(126)	99%(165)	96%(27)
	$_{345}$DTTPFGQQR$_{353}$	100%(87)	100%(126)	100%(165)	96%(27)
	$_{346}$TTPFGQQRV$_{354}$	100%(87)	100%(126)	100%(165)	96%(27)
	$_{347}$TPFGQQRVF$_{355}$	100%(87)	100%(126)	100%(157)	96%(27)
	$_{348}$PFGQQRVFK$_{356}$	100%(87)	100%(126)	100%(157)	96%(27)
	$_{349}$FGQQRVFKE$_{357}$	100%(87)	100%(126)	100%(157)	100%(27)
	$_{350}$GQQRVFKEK$_{358}$	100%(87)	100%(126)	99%(157)	100%(27)
	$_{354}$VFKEKVDTR$_{362}$	100%(87)	100%(126)	99%(157)	100%(27)
	$_{450}$CVYNMMGKREKKLGEFG$_{466}$				
	$_{450}$CVYNMMGKR$_{458}$	100%(87)	95%(126)	97%(155)	100%(27)
	$_{451}$VYNMMGKRE$_{459}$	100%(87)	95%(126)	97%(155)	100%(27)
	$_{452}$YNMMGKREK$_{460}$	100%(87)	97%(126)	97%(155)	100%(27)
	$_{453}$NMMGKREKK$_{461}$	100%(87)	97%(126)	97%(155)	100%(27)
	$_{454}$MMGKREKKL$_{462}$	100%(87)	94%(126)	97%(155)	100%(27)
	$_{457}$KREKKLGEF$_{465}$	100%(87)	97%(126)	99%(155)	100%(27)
	$_{458}$REKKLGEFG$_{466}$	100%(87)	97%(126)	99%(155)	100%(27)
	$_{468}$AKGSRAIWYMWLGAR$_{482}$				
	$_{469}$KGSRAIWYM$_{477}$	98%(87)	99%(126)	99%(155)	100%(27)
	$_{470}$GSRAIWYMW$_{478}$	98%(87)	100%(128)	99%(155)	100%(27)
	$_{471}$SRAIWYMWL$_{479}$	98%(87)	100%(128)	100%(155)	100%(27)
	$_{473}$AIWYMWLGA$_{481}$	98%(87)	100%(128)	99%(157)	100%(27)
	$_{474}$IWYMWLGAR$_{482}$	97%(87)	100%(128)	99%(157)	100%(27)
	$_{531}$YADDTAGWDTRIT$_{543}$				
	$_{531}$YADDTAGWD$_{539}$	100%(87)	100%(128)	99%(157)	100%(27)

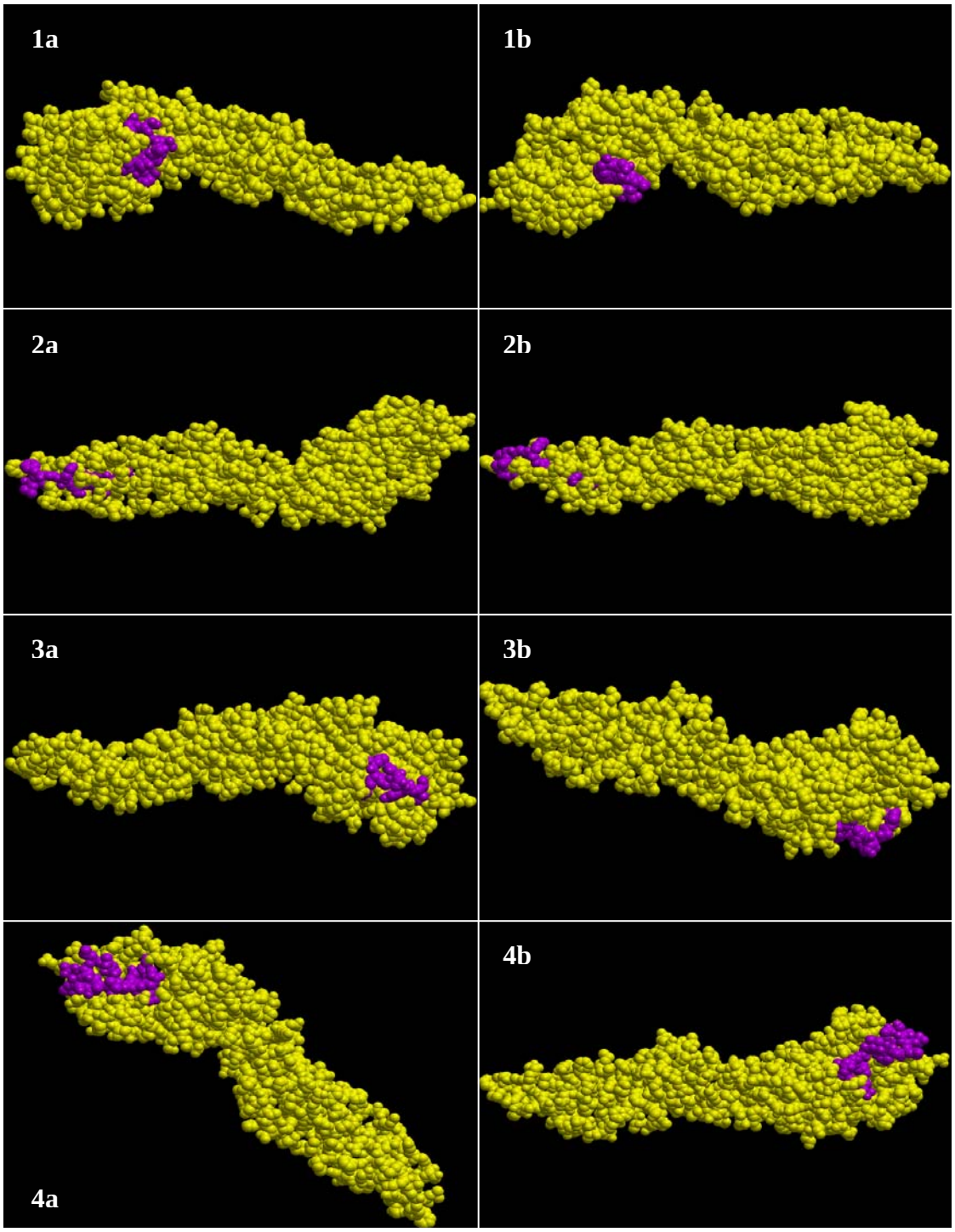
$_{534}$DTAGWDTRI$_{542}$	100%(87)	100%(128)	99%(157)	100%(27)
$_{568}$IFKLTYQNKVV$_{578}$				
$_{568}$IFKLTYQNK$_{576}$	100%(87)	98%(130)	97%(159)	100%(27)
$_{569}$FKLTYQNKV$_{577}$	100%(87)	97%(130)	98%(159)	100%(27)
$_{570}$KLTYQNKVV$_{578}$	100%(87)	97%(130)	98%(159)	100%(27)
$_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$				
$_{599}$RGSGQVGTY$_{607}$	100%(87)	88%(130)	98%(159)	100%(27)
$_{601}$SGQVGTYGL$_{609}$	100%(87)	88%(130)	98%(159)	100%(27)
$_{604}$VGTYGLNTF$_{612}$	100%(87)	88%(130)	97%(159)	100%(27)
$_{605}$GTYGLNTFT$_{613}$	100%(87)	88%(114)	99%(158)	100%(27)
$_{606}$TYGLNTFTN$_{614}$	100%(87)	98%(130)	99%(159)	100%(27)
$_{607}$YGLNTFTNM$_{615}$	98%(87)	98%(130)	99%(159)	100%(27)
$_{658}$RMAISGDDCVVKP$_{670}$				
$_{659}$MAISGDDCV$_{667}$	100%(87)	100%(130)	97%(159)	100%(27)
$_{660}$AISGDDCVV$_{668}$	100%(87)	100%(130)	97%(159)	100%(27)
$_{661}$ISGDDCVVK$_{669}$	100%(87)	100%(130)	97%(159)	100%(27)
$_{707}$VPFCSHHFH$_{715}$				
$_{707}$VPFCSHHFH$_{715}$	98%(87)	100%(130)	100%(159)	96%(27)
$_{765}$LMYFHRRDLRLA$_{776}$				
$_{765}$LMYFHRRDL$_{773}$	100%(87)	98%(130)	99%(159)	100%(27)
$_{766}$MYFHRRDLR$_{774}$	100%(87)	98%(130)	99%(159)	100%(27)
$_{767}$YFHRRDLRL$_{775}$	100%(87)	98%(130)	99%(159)	100%(27)
$_{768}$FHRRDLRLA$_{776}$	100%(87)	98%(130)	98%(159)	100%(27)
$_{790}$PTSRTTWSIHA$_{800}$				
$_{790}$PTSRTTWSI$_{798}$	99%(87)	98%(128)	98%(157)	100%(27)
$_{792}$SRTTWSIHA$_{800}$	99%(87)	98%(128)	100%(157)	100%(27)

[a] Amino acid positions of the pan-DENV sequences and the predicted nonamers are numbered according to the sequence alignments of the 4 DENV types

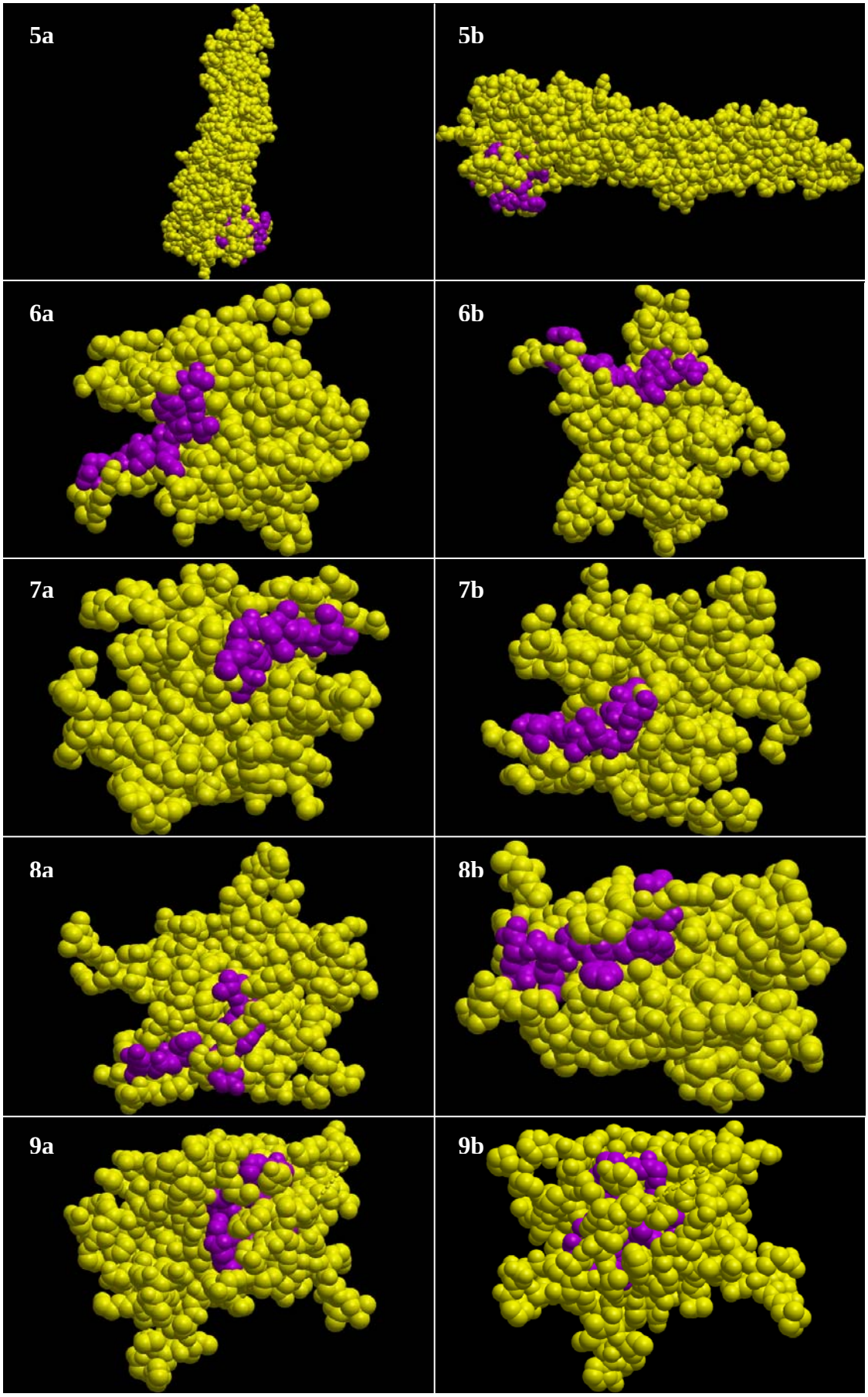
[b] Rounded to whole number

[c] The total number of sequences analyzed (2005 dataset) may not match the total number of sequences collected for each protein because both partial and full-length sequences were used for the sequence alignments, with the results that some regions have more sequence information than others

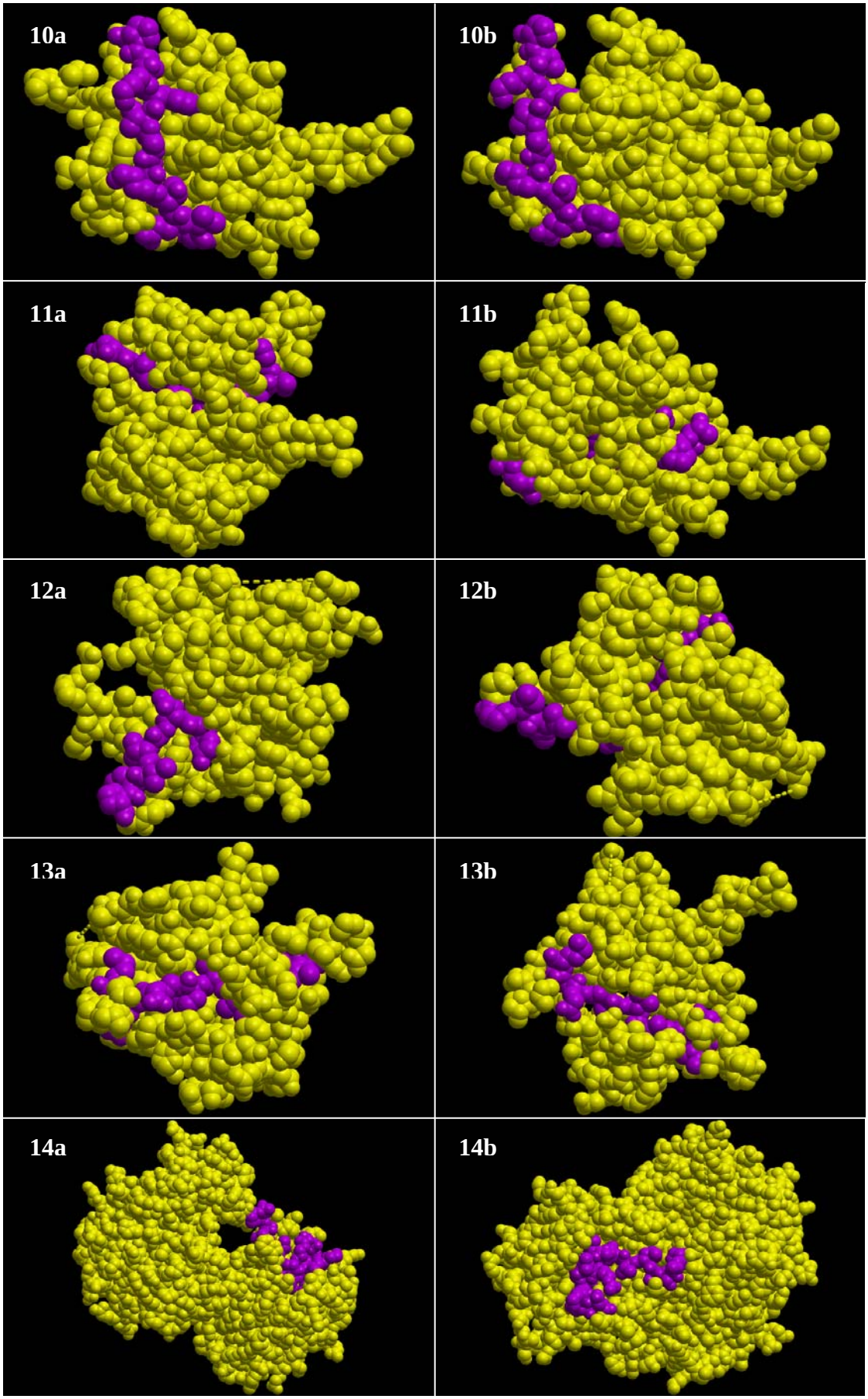
Appendix 6: The localization of pan-WNV sequences (in purple) on the three dimensional structure of the respective WNV proteins (E - 2HG0, NS3 - 2IJO and NS5 -2HFZ). Abbreviations: (E) major portion exposed, (P) partially exposed, (B) major portion buried.



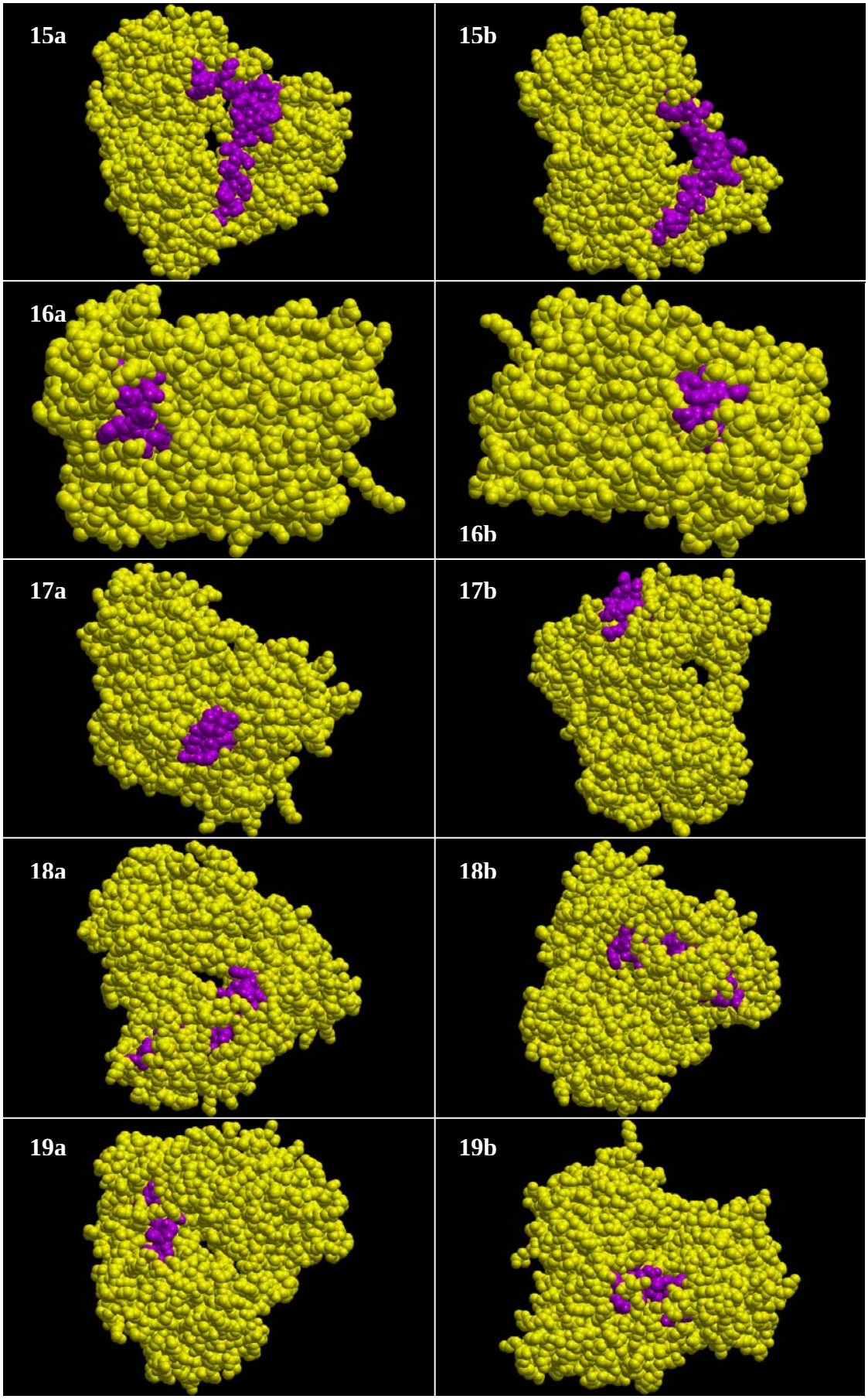
1. $E_{1\text{-}11}$ (E) | 2. $E_{104\text{-}117}$ (P) | 3. $E_{293\text{-}301}$ (E) | 4. $E_{338\text{-}356}$ (E)



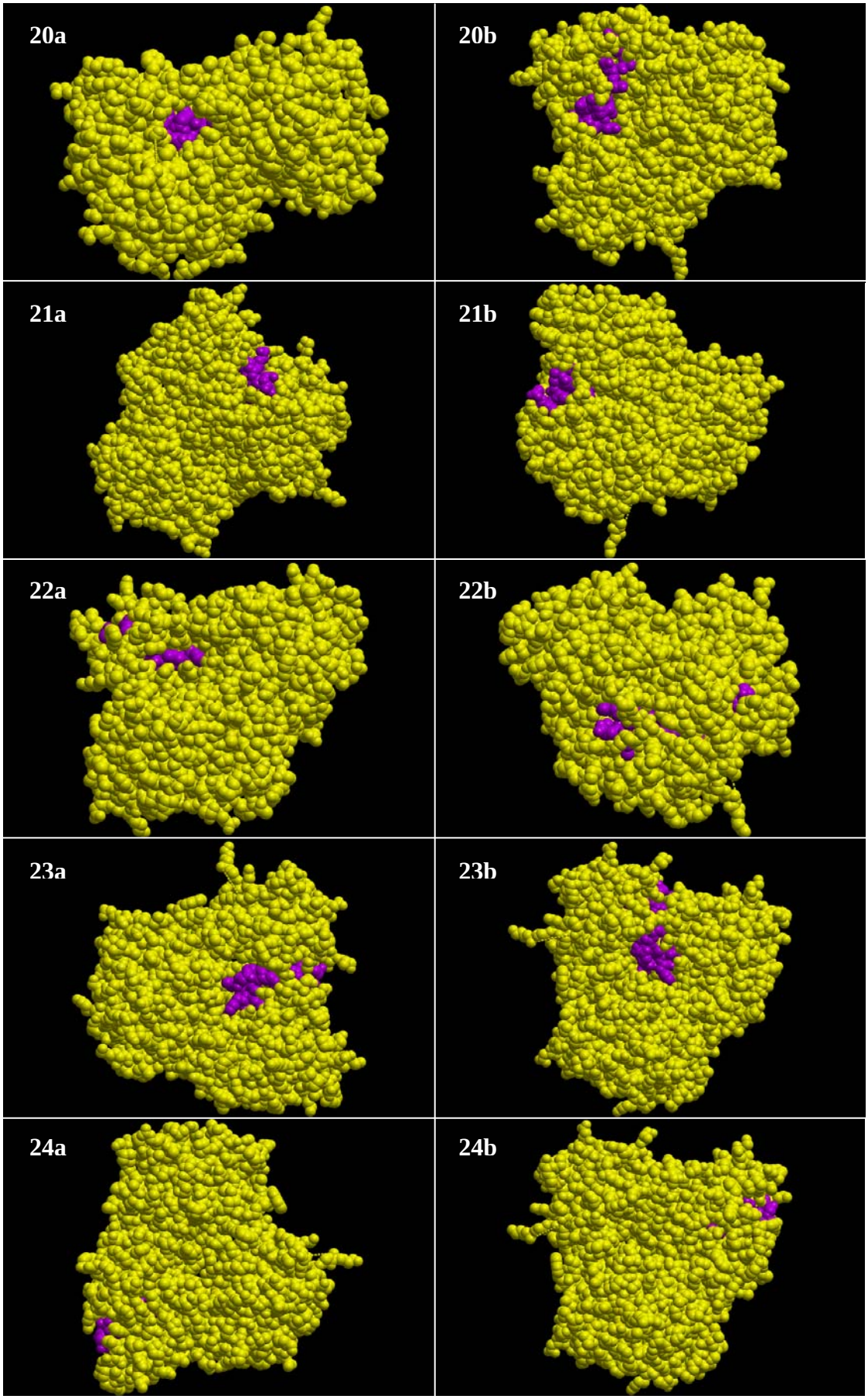
5. $E_{370-381}$ (P) | 6. $NS3_{20-29}$ (E) | 7. $NS3_{52-61}$ (E) | 8. $NS3_{63-72}$ (P) | 9. $NS3_{74-83}$ (B)



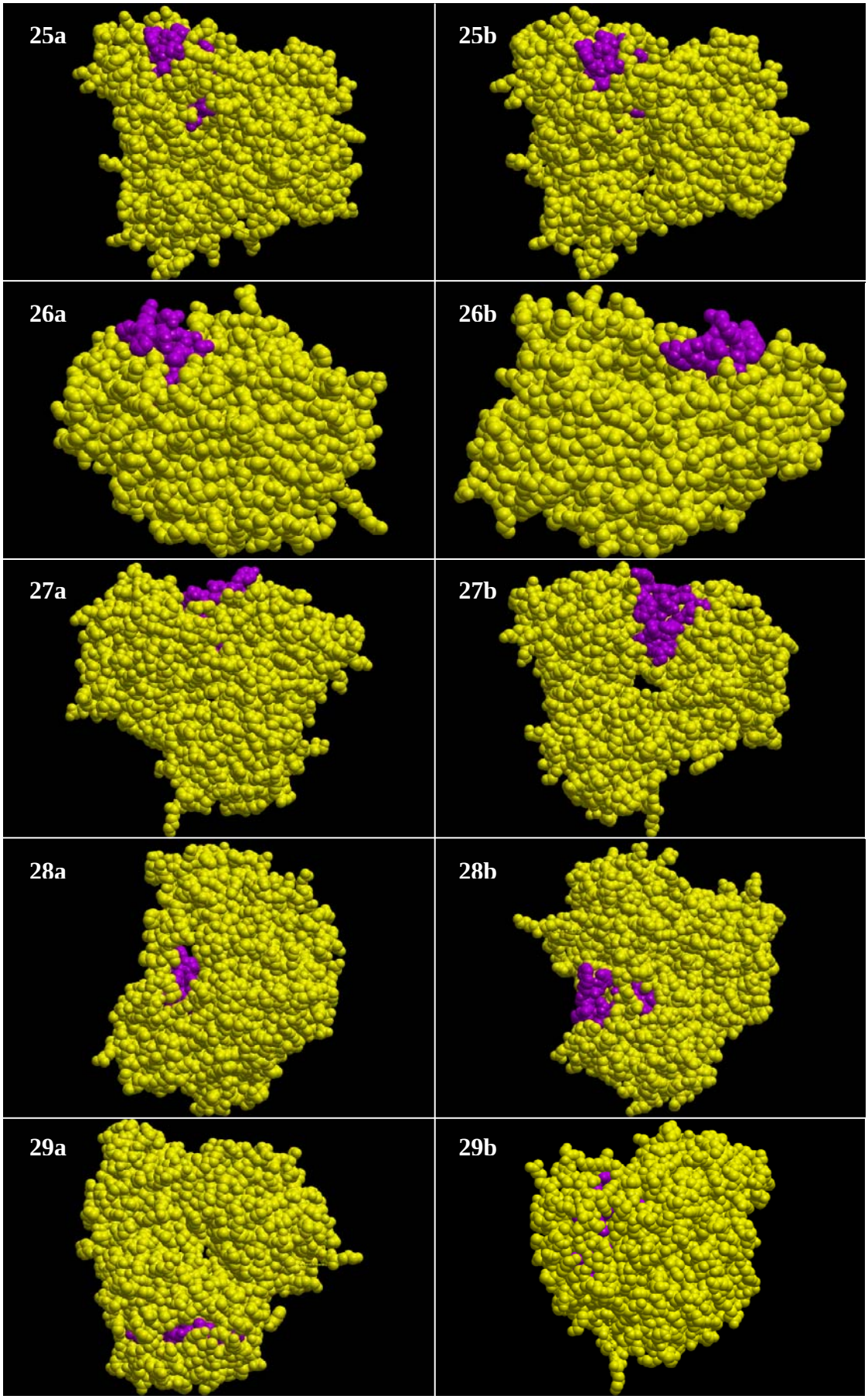
10. $NS3_{108-119}$ (E) | 11. $NS3_{131-142}$ (B) | 12. $NS3_{145-157}$ (P) | 13. $NS3_{161-171}$ (P) | 14. $NS5_{318-335}$ (E)



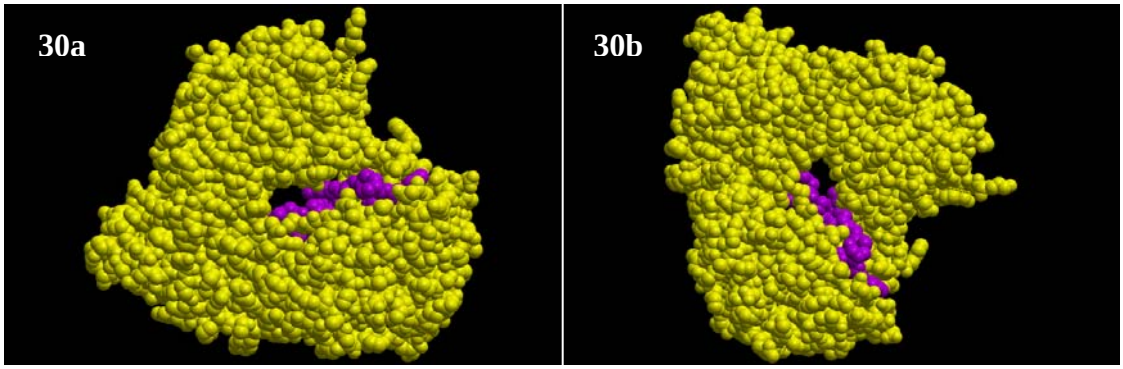
15. $NS5_{340-368}$ (E) | 16. $NS5_{375-384}$ (P) | 17. $NS5_{440-449}$ (E) | 18. $NS5_{472-500}$ (P) | 19. $NS5_{504-519}$ (P)



20. $NS5_{533-545}$ (P) | 21. $NS5_{548-557}$ (E) | 22. $NS5_{571-585}$ (B) | 23. $NS5_{596-618}$ (P) | 24. $NS5_{620-631}$ (B)



25. $NS5_{662-680}$ (P) | 26. $NS5_{689-702}$ (E) | 27. $NS5_{704-721}$ (E) | 28. $NS5_{741-767}$ (P) | 29. $NS5_{769-789}$ (B)



30. $NS5_{792-803}$ (P)

Appendix 7: Percentage representation of pan-WNV sequences in other flaviviruses.

WNV protein	Pan-WNV sequence	Species (#) [a]	Percentage representation (%) | Total number of sequences analyzed [b]							
			DENV	JEV	LIV	OMSK	PV	LEV	TBEV	YFV
prM	125-ESWILRNPGYALVA-138	5	25|524					100|29		
	158-LLLLVAPAYS-167	7		98|100						
E	104-GCGLFGKGSIDTCA-117	31	52|1295	99|245	100|15			100|97	98|150	77|163
	293-LKGTTYGVC-301	1		1|256						
	370-ELEPPFGDSYIV-381	11	40|1402	98|253				94|77		
	417-LGDTAWDFGS-426	9	84|1296	96|252				100|77		
	449-LFGGMSWITQGL-460	5		99|244				100|77		
NS1	114-GWKAWGKSI-122	2		95|58						
	195-HSDLSYWIES-204	4		95|57				100|27		
	209-TWKLERAVLGEVKSCTWPETHTLWG-233	6		100|57				100|27		
	276-DFDYCPGTTVT-286	4		2|58				96|27		
	313-CRSCTLPPLR-322	6	92|335	2|58						
	328-GCWYGMEIRP-337	10	97|329	98|58				100|27		
NS2a	4-DMIDPFQLGL-13	3		12|58						
NS2b	12-GLMFAIVGGLAELD-25	3		100|55						
NS3	1-GGVLWDTPSP-10	1	32|247							
	145-DVIGLYGNGVIMP-157	4	9|258							
	191-VLDLHPGAGKTR-202	11	40|255					100|26		
	235-ALRGLPIRY-243	2	40|255							
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282	25		100|53				100|26	100|17	5|22
	288-HFTDPASIAARGYI-301	12	80|245	98|53				100|26		100|22
	310-AAAIFMTATPPG-321	11	100|245	100|53				96|26		
	357-GKTVWFVPSV-366	8	99|275							

WNV protein	Pan-WNV sequence	Species (#) [a]	Percentage representation (%) | Total number of sequences analyzed [b]							
			DENV	JEV	LIV	OMSK	PV	LEV	TBEV	YFV
	408-TTDISEMGANF-418	34	99|275	100|53				100|26	100|17	
	451-TAASAAQRRGR-461	29	72|273	100|53				100|26	94|17	
	526-LRGEERKNFLE-536	2		2|53				100|26		
	540-TADLPVWLA-548	3		100|53						
	563-WCFDGPRTNT-572	1		100|53						
NS4a	43-ALEELPDALQT-53	3		100|52						
	115-MIVLIPEPEKQRSQTDNQLA-134	10	35|239	100|52				100|26		
NS4b	138-AQRRTAAGIMKN-149	10	69|248	100|52				100|26		
	156-VATDVPELER-165	3		100|52						
NS5	79-DLGCGRGGWCYYMATQK-95	36	99|246	94|52			88|17	100|27	100|31	100|21
	107-GPGHEEPQLVQSYGWNIVTMKS-128	6						96|27		
	141-DTLLCDIGES-150	13	99|245	2|52						100|21
	208-RNPLSRNSTHEMYWVS-223	30	99|244				100|17	96|27	100|17	100|21
	235-MTSQVLLGRMEK-246	1		100|52						
	259-NLGSGTRAVG-268	5		100|52						
	299-NHPYRTWNYHGSY-311	5						100|26		
	318-SASSLVNGVVRLLSKPWD-335	6		100|52				100|26		
	340-VTTMAMTDTTPFGQQRVFKEKVDTKAPEP-368	30	100|306	100|52			100|27	100|26	65|17	100|21
	375-VLNETTNWLW-384	1		100|52						
	404-KVNSNAALGAMFEEQNQW-421	6		92|52				96|26		
	451-TCIYNMMGKREK-462	34	99|299	98|52			84|19	96|26	88|17	100|21
	472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500	55	99|304	100|52				100|29	100|18	100|22
	504-NSGGGVEGLGLQKLGY-519	9		2|52						
	533-YADDTAGWDTRIT-545	59	100|300	98|52		100|13	94|17	100|27	100|18	100|22
	548-DLENEAKVLE-557	2		100|52						

WNV protein	Pan-WNV sequence	Species (#) [a]	Percentage representation (%) | Total number of sequences analyzed [b]							
			DENV	JEV	LIV	OMSK	PV	LEV	TBEV	YFV
	571-IELTYRHKVVKVMRP-585	10		98|52				100|27		
	596-ISREDQRGSGQVVTYALNTFTNL-618	61	3|302	100|52		100|13	94|17	100|27	100|18	100|22
	662-RMAVSGDDCVVKPLDDRFA-680	30	98|303	100|52			94|17	100|27		95|22
	689-MSKVRKDIQEWKPS-702	9		96|52				96|27		
	704-GWYDWQQVPFCSNHFTEL-721	7	68|303	98|52						
	741-GRARISPGAGWNVRDTACLAKSYAQMW-767	9		98|53				100|27		
	769-LLYFHRRDLRLMANAICSAVP-789	45	100|302	100|53				96|28		
	792-WVPTGRTTWSIH-803	38						100|27	92|24	

[a] The species column indicates the total number of viral species that share the pan-WNV sequence

[b] Percentage representation of WNV sequences in other viral species is only shown for species with at least a total of 10 sequences reported at NCBI Entrez protein database. These viral species include: DENV, *Dengue virus type 1, 2, 3 or 4*; JEV, *Japanese encephalitis virus*; LIV, *Louping ill virus*; OMSK, *Omsk hemorrhagic fever virus*; PV, *Powassan virus*; LEV, *St. Louis encephalitis virus*; TBEV, *Tick-born encephalitis*; and YFV, *Yellow fever virus*. However, despite having a total of ≥ 10 sequences reported, some of these viruses had less than 10 of the relevant conserved sequence (indicated by cells shaded in grey). Empty cells indicate no match between the pan-WNV sequences and the *Flavivirus*.

Appendix 8: Putative HLA supertype-restricted binding nonamer peptides in pan-WNV sequences, predicted by immunoinformatics algorithms (NetCTL, Multipred (MP), ARB and TEPITOPE (TP)).

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
prM	125-ESWILRNPGYALVA-138																				
	126-SWILRNPGY-134	A1			A24	A26							B62								
	127-WILRNPGYA-135																			DR	
	128-ILRNPGYAL-136		A2				B7	B8					B62			A2				DR	
	129-LRNPGYALV-137								B27							A2				DR	
	130-RNPGYALVA-138															A2					
	158-LLLLVAPAYS-167																				
	158-LLLLVAPAY-166	A1		A3		A26		B8				B58	B62							DR	DR
	159-LLLVAPAYS-167															A2				DR	DR
E	1-FNCLGMSNRDF-11																				
	1-FNCLGMSNR-9																A3			DR	
	104-GCGLFGKGSIDTCA-117																				
	107-LFGKGSIDT-115																			DR	
	293-LKGTTYGVC-301																				
	293-LKGTTYGVC-301																			DR	
	338-SVASLNDLTPVGRLVTVNP-356																				
	338-SVASLNDLT-346															A2					
	340-ASLNDLTPV-348		A2											A2		A2					
	346-TPVGRLVTV-354						B7	B8						A2		A2					
	370-ELEPPFGDSYIV-381																				
	371-LEPPFGDSY-379	A1									B44		B62								
	449-LFGGMSWITQGL-460																				
	449-LFGGMSWIT-457																			DR	
	450-FGGMSWITQ-458																				DR
	452-GMSWITQGL-460		A2					B8					B62			A2					
NS1	58-RSVSRLEHQMW-68																				
	59-SVSRLEHQM-67					A26						B58	B62								

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL											MULTIPRED			ARB			MP	TP	
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	60-VSRLEHQMW-68											B58									
	154-EVEDFGFGL-162																				
	154-EVEDFGFGL-162	A1				A26				B39											
	195-HSDLSYWIES-204																				
	195-HSDLSYWIE-203	A1																			
	196-SDLSYWIES-204																		B44		
	209-TWKLERAVLGEVKSCTWPETHTLWG-233																				
	209-TWKLERAVL-217				A24			B8		B39											
	210-WKLERAVLG-218																			DR	
	211-KLERAVLGE-219													A2							
	212-LERAVLGEV-220										B44										
	213-ERAVLGEVK-221								B27												
	215-AVLGEVKSC-223													A2							
	216-VLGEVKSCT-224													A2							
	217-LGEVKSCTW-225											B58									
	218-GEVKSCTWP-226										B44										
	220-VKSCTWPET-228																			DR	
	223-CTWPETHTL-231	A1	A2							B39		B58		A2							
	224-TWPETHTLW-232				A24							B58									
	276-DFDYCPGTTVT-286																				
	277-FDYCPGTTV-285															A2			B44	DR	
	313-CRSCTLPPLR-322																				
	313-CRSCTLPPL-321								B27	B39											
	314-RSCTLPPLR-322			A3													A3				
	328-GCWYGMEIRP-337																				
	328-GCWYGMEIR-336																A3				
NS2a	4-DMIDPFQLGL-13																				
	5-MIDPFQLGL-13	A1	A2							B39		B58		A2		A2					
	69-NSGGDVVHLALMATF-83																				
	69-NSGGDVVHL-77									B39											
	71-GGDVVHLAL-79									B39											

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED				ARB		MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	72-GDVVHLALM-80																		B44		
	74-VVHLALMAT-82																			DR	DR
	75-VHLALMATF-83				A24			B8	B27	B39										DR	
NS2b	1-GWPATEVMTA-10																				
	2-WPATEVMTA-10						B7											B7		DR	
	12-GLMFAIVGGLAELD-25																				
	13-LMFAIVGGL-21		A2						B27				B62	A2		A2				DR	DR
	14-MFAIVGGLA-22																			DR	
	15-FAIVGGLAE-23																			DR	DR
	16-AIVGGLAEL-24		A2			A26	B7						B62	A2		A2					
	17-IVGGLAELD-25																			DR	
	32-PMTIAGLMF-40																				
	32-PMTIAGLMF-40	A1			A24							B58	B62								
	108-SAYTPWAILPS-118																				
	108-SAYTPWAIL-116						B7			B39		B58	B62								
	110-YTPWAILPS-118															A2				DR	DR
NS3	52-TTKGAALMSG-61																				
	52-TTKGAALMS-60														A3						
	63-GRLDPYWGSV-72																				
	63-GRLDPYWGS-71								B27												
	64-RLDPYWGSV-72	A1	A2													A2					
	74-EDRLCYGGPW-83																				
	75-DRLCYGGPW-83								B27												
	108-NVQTKPGVFKTP-119																				
	108-NVQTKPGVF-116	A1			A24		B7	B8					B62								
	109-VQTKPGVFK-117			A3											A3					DR	
	110-QTKPGVFKT-118														A3						
	131-PTGTSGSPIVDK-142																				
	134-TSGSPIVDK-142			A3											A3						
	145-DVIGLYGNGVIMP-157																				
	146-VIGLYGNGV-154															A2				DR	

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	147-IGLYGNGVI-155																			DR	
	148-GLYGNGVIM-156		A2										B62								
	149-LYGNGVIMP-157																			DR	
	161-YISAIVQGERM-171																				
	161-YISAIVQGE-169													A2						DR	
	162-ISAIVQGER-170			A3											A3		A3			DR	
	163-SAIVQGERM-171					A26						B58									
	235-ALRGLPIRY-243																				
	235-ALRGLPIRY-243	A1		A3		A26							B62		A3						
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282																				
	256-EIVDVMCHA-264					A26								A2							
	257-IVDVMCHAT-265																			DR	
	258-VDVMCHATL-266							B8			B44								B44		
	259-DVMCHATLT-267														A3	A2					
	260-VMCHATLTH-268			A3									B62	A2			A3			DR	DR
	261-MCHATLTHR-269			A3												A2	A3			DR	
	262-CHATLTHRL-270									B39											
	263-HATLTHRLM-271						B7	B8													
	264-ATLTHRLMS-272														A3						
	265-TLTHRLMSP-273													A2							
	266-LTHRLMSPH-274			A3													A3			DR	
	268-HRLMSPHRV-276								B27	B39				A2							
	270-LMSPHRVPN-278																			DR	
	271-MSPHRVPNY-279	A1		A3		A26						B58	B62								
	273-PHRVPNYNL-281									B39											
	274-HRVPNYNLF-282	A1			A24			B8	B27	B39											
	288-HFTDPASIAARGYI-301																				
	289-FTDPASIAA-297	A1	A2													A2					
	292-PASIAARGY-300	A1																			
	293-ASIAARGYI-301																A3				
	310-AAAIFMTATPPG-321																				

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	310-AAAIFMTAT-318						B7									A2					
	313-IFMTATPPG-321																			DR	DR
	337-QTEIPDRAWN-346																				
	337-QTEIPDRAW-345											B58									
	357-GKTVWFVPSV-366																				
	358-KTVWFVPSV-366		A2			A26						B58		A2		A2					
	385-QLNRKSYETEYPKCKN-400																				
	385-QLNRKSYET-393													A2	A3						
	387-NRKSYETEY-395	A1						B8	B27												
	389-KSYETEYPK-397			A3					B27						A3		A3				
	391-YETEYPKCK-399										B44										
	408-TTDISEMGANF-418																				
	408-TTDISEMGA-416	A1																			
	410-DISEMGANF-418	A1				A26							B62								
	422-RVIDSRKSVKP-432																				
	422-RVIDSRKSV-430						B7						B62	A2							
	423-VIDSRKSVK-431			A3											A3						
	451-TAASAAQRRGR-461																				
	451-TAASAAQRR-459			A3											A3		A3				
	453-ASAAQRRGR-461														A3		A3				
	526-LRGEERKNFLE-536																				
	526-LRGEERKNF-534								B27												
	527-RGEERKNFL-535							B8													
	540-TADLPVWLA-548																				
	540-TADLPVWLA-548	A1														A2					
	563-WCFDGPRTNT-572																				
	563-WCFDGPRTN-571																			DR	
NS4a	19-KTWEALDTMYVVATA-33																				
	19-KTWEALDTM-27		A2			A26						B58		A2		A2					
	20-TWEALDTMY-28	A1																			
	21-WEALDTMYV-29										B44					A2			B44	DR	DR

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	22-EALDTMYVV-30		A2			A26								A2							
	23-ALDTMYVVA-31	A1	A2											A2		A2					
	24-LDTMYVVAT-32																		B44	DR	
	25-DTMYVVATA-33					A26										A2					
	43-ALEELPDALQT-53																				
	43-ALEELPDAL-51		A2							B39				A2							
	45-EELPDALQT-53										B44										
	101-GTKIAGMLLLSLL-113																				
	101-GTKIAGMLL-109	A1																			
	102-TKIAGMLLL-110					A26				B39				A2							
	103-KIAGMLLLS-111			A3										A2		A2					
	104-IAGMLLLSL-112		A2				B7													DR	DR
	115-MIVLIPEPEKQRSQTDNQLA-134																				
	115-MIVLIPEPE-123																			DR	
	116-IVLIPEPEK-124			A3											A3		A3				
	118-LIPEPEKQR-126														A3		A3				
	125-QRSQTDNQL-133								B27	B39											
NS4b	68-TSLTSINVQASAL-80																				
	69-SLTSINVQA-77		A2											A2		A2					
	70-LTSINVQAS-78																			DR	DR
	71-TSINVQASA-79					A26										A2					
	72-SINVQASAL-80						B7	B8					B62	A2							
	138-AQRRTAAGIMKN-149																				
	138-AQRRTAAGI-146												B62	A2							
	139-QRRTAAGIM-147								B27												
	140-RRTAAGIMK-148								B27						A3						
	141-RTAAGIMKN-149			A3																	
	156-VATDVPELER-165																				
	156-VATDVPELE-164																			DR	
	157-ATDVPELER-165	A1													A3		A3				
	208-VTLWENGASSVWNATTAIGLCH-229																				

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	209-TLWENGASS-217													A2		A2					
	210-LWENGASSV-218													A2							
	211-WENGASSVW-219										B44	B58	B62								
	215-ASSVWNATT-223	A1																			
	217-SVWNATTAI-225		A2			A26	B7						B62	A2		A2					
	219-WNATTAIGL-227															A2				DR	
	221-ATTAIGLCH-229	A1		A3											A3		A3				
NS5	60-AKLRWLVER-68																				
	60-AKLRWLVER-68								B27						A3						
	79-DLGCGRGGWCYYMATQK-95																				
	81-GCGRGGWCY-89	A1																			
	82-CGRGGWCYY-90	A1				A26							B62								
	83-GRGGWCYYM-91								B27												
	87-WCYYMATQK-95			A3											A3		A3			DR	
	107-GPGHEEPQLVQSYGWNIVTMKS-128																				
	107-GPGHEEPQL-115						B7														
	111-EEPQLVQSY-119					A26					B44										
	115-LVQSYGWNI-123		A2											A2		A2				DR	
	116-VQSYGWNIV-124												B62			A2				DR	
	118-SYGWNIVTM-126				A24					B39											
	119-YGWNIVTMK-127								B27												
	141-DTLLCDIGES-150																				
	142-TLLCDIGES-150															A2					
	152-SSAEVEEHRT-161																				
	152-SSAEVEEHR-160														A3		A3				
	168-VEDWLHRGP-176																				
	168-VEDWLHRGP-176										B44										
	208-RNPLSRNSTHEMYWVS-223																				
	211-LSRNSTHEM-219						B7					B58	B62					B7		DR	
	212-SRNSTHEMY-220	A1						B8	B27												
	213-RNSTHEMYW-221											B58									

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																				
		Class I																			Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	235-MTSQVLLGRMEK-246																				
	235-MTSQVLLGR-243	A1		A3										A2	A3	A2	A3				
	238-QVLLGRMEK-246			A3											A3		A3				
	259-NLGSGTRAVG-268																				
	259-NLGSGTRAV-267		A2											A2		A2					
	299-NHPYRTWNYHGSY-311																				
	299-NHPYRTWNY-307	A1																			
	302-YRTWNYHGS-310								B27											DR	
	303-RTWNYHGSY-311	A1		A3		A26		B8	B27			B58	B62		A3		A3				
	318-SASSLVNGVVRLLSKPWD-335																				
	318-SASSLVNGV-326		A2											A2		A2					
	319-ASSLVNGVV-327	A1																			
	320-SSLVNGVVR-328														A3		A3				
	321-SLVNGVVRL-329	A1	A2			A26				B39			B62	A2		A2					
	322-LVNGVVRLL-330		A2										B62	A2		A2				DR	
	323-VNGVVRLLS-331																			DR	DR
	326-VVRLLSKPW-334						B7						B62							DR	DR
	327-VRLLSKPWD-335																			DR	DR
	340-VTTMAMTDTTPFGQQRVFKEKVDTKAPEP-368																				
	340-VTTMAMTDT-348																			DR	
	343-MAMTDTTPF-351	A1			A24		B7	B8				B58	B62					B7		DR	
	345-MTDTTPFGQ-353	A1																			
	347-DTTPFGQQR-355														A3		A3				
	348-TTPFGQQRV-356	A1				A26															
	349-TPFGQQRVF-357						B7	B8					B62					B7			
	350-PFGQQRVFK-358														A3						
	351-FGQQRVFKE-359																			DR	
	352-GQQRVFKEK-360			A3					B27						A3						
	356-VFKEKVDTK-364														A3						
	375-VLNETTNWLW-384																				
	375-VLNETTNWL-383		A2							B39			B62	A2		A2					

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED		ARB			MP	TP	
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	376-LNETTNWLW-384	A1										B58									
	404-KVNSNAALGAMFEEQNQW-421																				
	404-KVNSNAALG-412														A3						
	405-VNSNAALGA-413															A2				DR	DR
	406-NSNAALGAM-414					A26															
	407-SNAALGAMF-415					A26							B62								
	410-ALGAMFEEQ-418													A2							
	413-AMFEEQNQW-421											B58	B62	A2	A3						
	440-EREAHLRGEC-449																				
	441-REAHLRGEC-449										B44										
	451-TCIYNMMGKREK-462																				
	451-TCIYNMMGK-459					A26									A3		A3				
	452-CIYNMMGKR-460			A3											A3		A3				
	453-IYNMMGKRE-461																			DR	
	454-YNMMGKREK-462														A3		A3			DR	DR
	472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500																				
	472-GSRAIWFMW-480											B58									
	473-SRAIWFMWL-481								B27	B39											
	475-AIWFMWLGA-483		A2													A2					
	476-IWFMWLGAR-484																A3			DR	
	477-WFMWLGARF-485				A24			B8					B62							DR	
	478-FMWLGARFL-486		A2					B8					B62			A2				DR	DR
	480-WLGARFLEF-488	A1			A24			B8					B62	A2						DR	
	483-ARFLEFEAL-491							B8	B27	B39	B44										
	485-FLEFEALGF-493	A1											B62							DR	
	486-LEFEALGFL-494										B44					A2					
	488-FEALGFLNE-496										B44									DR	
	490-ALGFLNEDH-498																A3				
	491-LGFLNEDHW-499											B58									
	492-GFLNEDHWL-500				A24						B44										
	504-NSGGGVEGLGLQKLGY-519																				

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED		ARB				MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	504-NSGGGVEGL-512									B39											
	508-GVEGLGLQK-516			A3											A3						
	509-VEGLGLQKL-517										B44										
	511-GLGLQKLGY-519	A1		A3									B62								
	533-YADDTAGWDTRIT-545																				
	533-YADDTAGWD-541																			DR	
	536-DTAGWDTRI-544	A1				A26								A2							
	548-DLENEAKVLE-557																				
	548-DLENEAKVL-556													A2							
	571-IELTYRHKVVKVMRP-585																				
	571-IELTYRHKV-579										B44								B44	DR	
	572-ELTYRHKVV-580							B8						A2							
	573-LTYRHKVVK-581			A3											A3		A3			DR	
	574-TYRHKVVKV-582				A24									A2							
	575-YRHKVVKVM-583							B8	B27	B39										DR	DR
	576-RHKVVKVMR-584														A3						
	596-ISREDQRGSGQVVTYALNTFTNL-618																				
	600-DQRGSGQVV-608												B62								
	602-RGSGQVVTY-610	A1		A3					B27			B58	B62								
	604-SGQVVTYAL-612						B7	B8		B39											
	606-QVVTYALNT-614														A3						
	607-VVTYALNTF-615	A1			A24	A26	B7					B58	B62							DR	
	608-VTYALNTFT-616															A2				DR	
	609-TYALNTFTN-617				A24																
	610-YALNTFTNL-618	A1	A2		A24	A26	B7	B8		B39	B44			A2		A2				DR	
	620-VQLVRMMEGEGV-631																				
	620-VQLVRMMEG-628																			DR	DR
	622-LVRMMEGEG-630																			DR	
	623-VRMMEGEGV-631																			DR	DR
	662-RMAVSGDDCVVKPLDDRFA-680																				
	663-MAVSGDDCV-671															A2		B7			

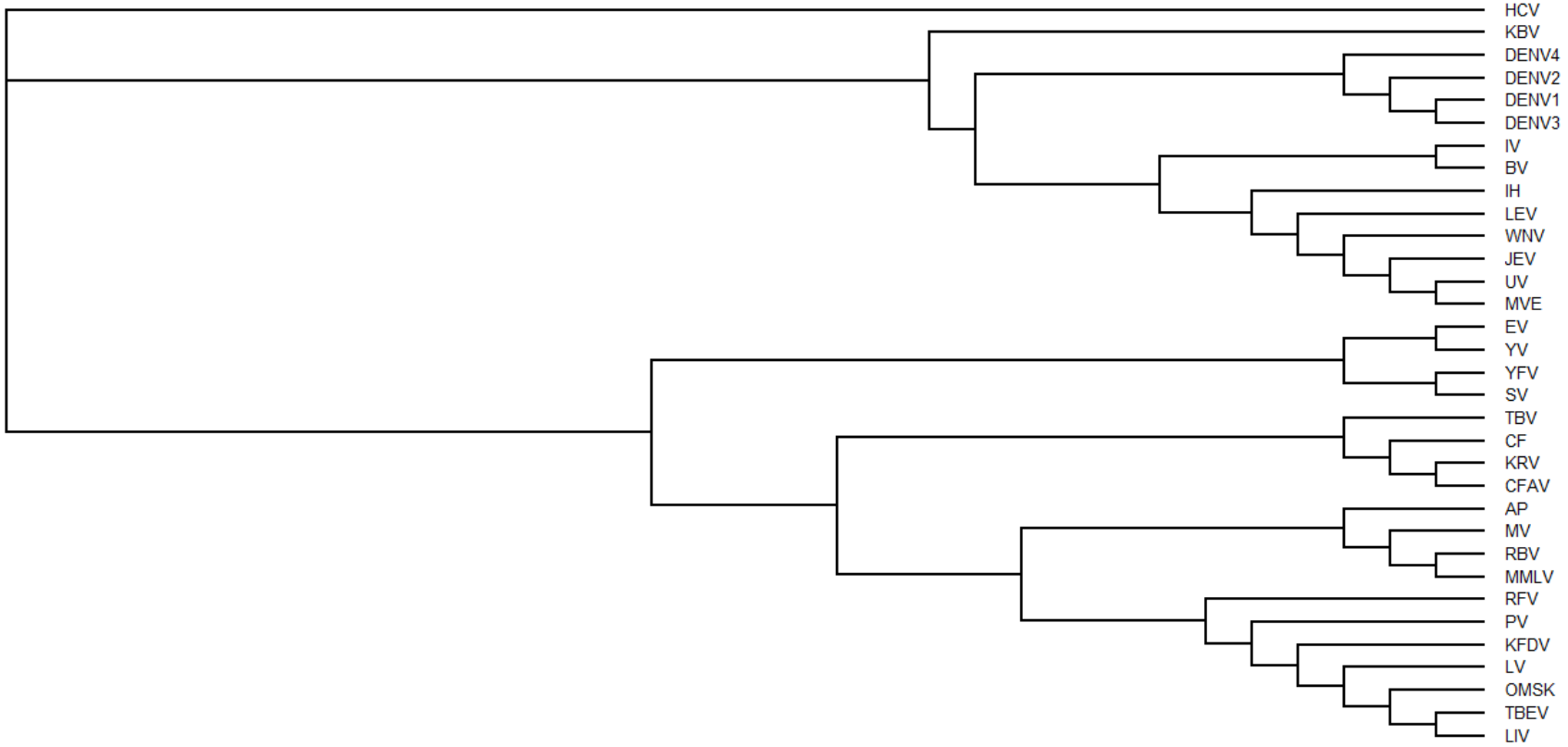
WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	664-AVSGDDCVV-672													A2		A2					
	667-GDDCVVKPL-675										B44								B44		
	670-CVVKPLDDR-678														A3						
	671-VVKPLDDRF-679												B62								
	689-MSKVRKDIQEWKPS-702																				
	691-KVRKDIQEW-699											B58	B62								
	692-VRKDIQEWK-700								B27						A3					DR	
	694-KDIQEWKPS-702																		B44		
	704-GWYDWQQVPFCSNHFTEL-721																				
	705-WYDWQQVPF-713	A1			A24			B8		B39										DR	
	708-WQQVPFCSN-716																			DR	
	709-QQVPFCSNH-717												B62				A3				
	710-QVPFCSNHF-718	A1			A24	A26							B62								
	711-VPFCSNHFT-719																	B7		DR	
	713-FCSNHFTEL-721	A1						B8		B39	B44			A2		A2					
	741-GRARISPGAGWNVRDTACLAKSYAQMW-767																				
	741-GRARISPGA-749								B27												
	742-RARISPGAG-750						B7														
	743-ARISPGAGW-751								B27												
	745-ISPGAGWNV-753	A1														A2					
	751-WNVRDTACL-759									B39						A2				DR	
	752-NVRDTACLA-760													A2		A2					
	753-VRDTACLAK-761	A1							B27												
	754-RDTACLAKS-762																		B44		
	755-DTACLAKSY-763	A1				A26							B62								
	758-CLAKSYAQM-766		A2			A26		B8						A2							
	759-LAKSYAQMW-767											B58								DR	
	769-LLYFHRRDLRLMANAICSAVP-789																				
	769-LLYFHRRDL-777							B8												DR	DR
	770-LYFHRRDLR-778														A3					DR	
	771-YFHRRDLRL-779				A24			B8		B39				A2						DR	

WNV Protein	Pan-WNV Sequence	HLA Supertype-Restriction of Predicted Nonamer Peptide [a]																			
		Class I																		Class II	
		NetCTL												MULTIPRED			ARB			MP	TP
		A1	A2	A3	A24	A26	B7	B8	B27	B39	B44	B58	B62	A2	A3	A2	A3	B7	B44	DR	DR
	772-FHRRDLRLM-780							B8												DR	
	773-HRRDLRLMA-781								B27												
	774-RRDLRLMAN-782								B27												
	776-DLRLMANAI-784							B8						A2							
	777-LRLMANAIC-785																			DR	DR
	778-RLMANAICS-786			A3										A2	A3						
	779-LMANAICSA-787		A2										B62	A2		A2				DR	DR
	780-MANAICSAV-788		A2											A2		A2					
	792-WVPTGRTTWSIH-803																				
	792-WVPTGRTTW-800						B7					B58	B62								
	793-VPTGRTTWS-801						B7														
	794-PTGRTTWSI-802				A24																

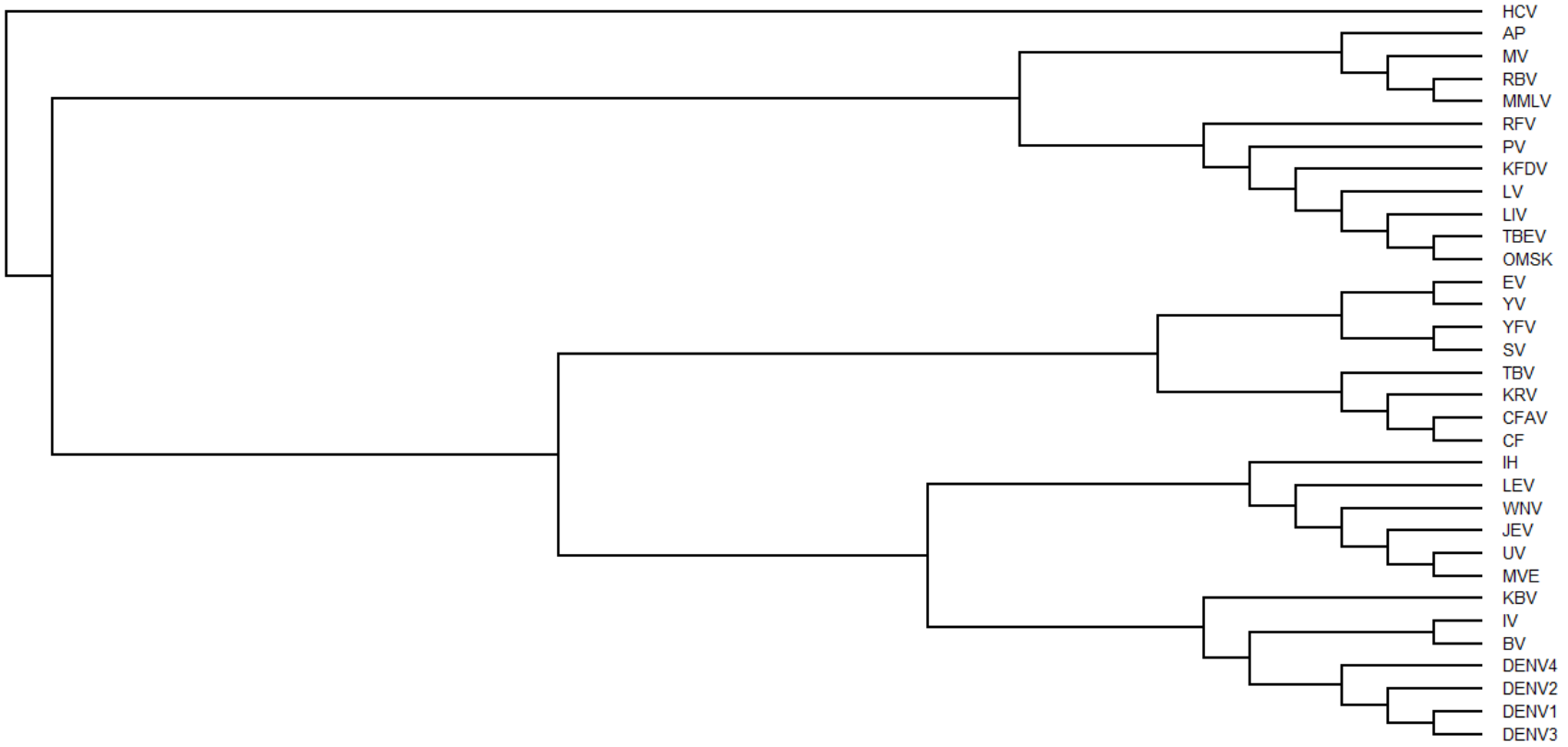
[a] Putative supertypes-restrictions of nonamer sequences with concurring predictions from at least two prediction tools are highlighted in grey. Hotspots with at least three sequential nonamers overlapping by eight amino acids, found in 7 of the 78 pan-WNV sequences, are each indicated by a box.

Appendix 9: Phylogenetic relationship of (A) polyprotein proteome sequences of selected 29 flaviviruses and B) sequences in the proteins of these flaviviruses that corresponded to 41 of the 44 pan-DENV sequences. The evolutionary relationship of the proteins containing these pan-DENV sequences is also provided in panel B. The pan-DENV sequences $_{296}$AARGYISTRV$_{305}$ (NS3) and $_{35}$PASAWTLYAVATT$_{47}$ (NS4b) were excluded from the analysis because of technical difficulty in generating their trees using the alignment data extracted from the flaviviruses. In addition, the pan-DENV sequence $_{383}$VIQLSRKTFD$_{392}$ (NS3) was ignored because the corresponding sequence in the DENV1 strain was a variant

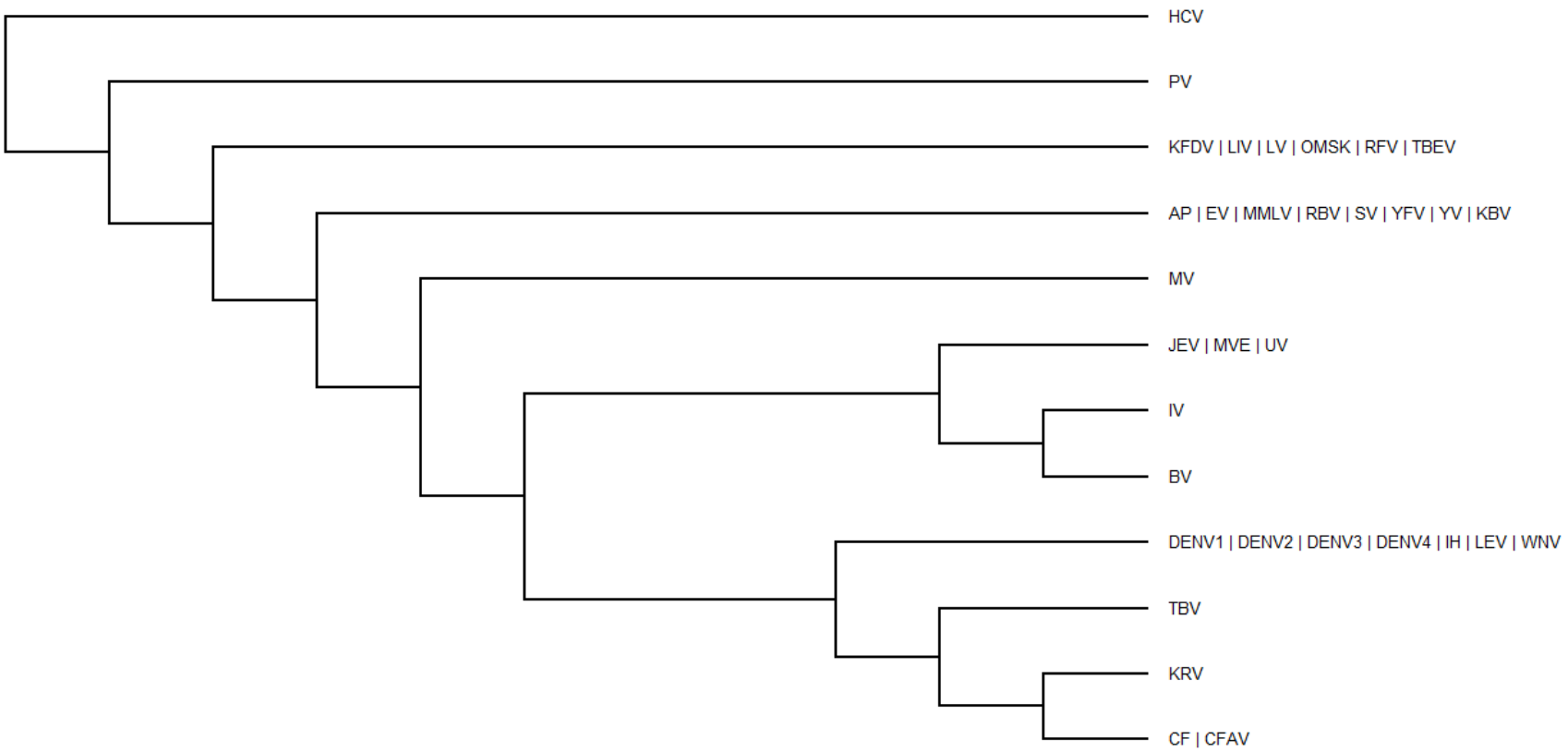
A) Evolution of full polyprotein proteome of selected *Flaviviruses*



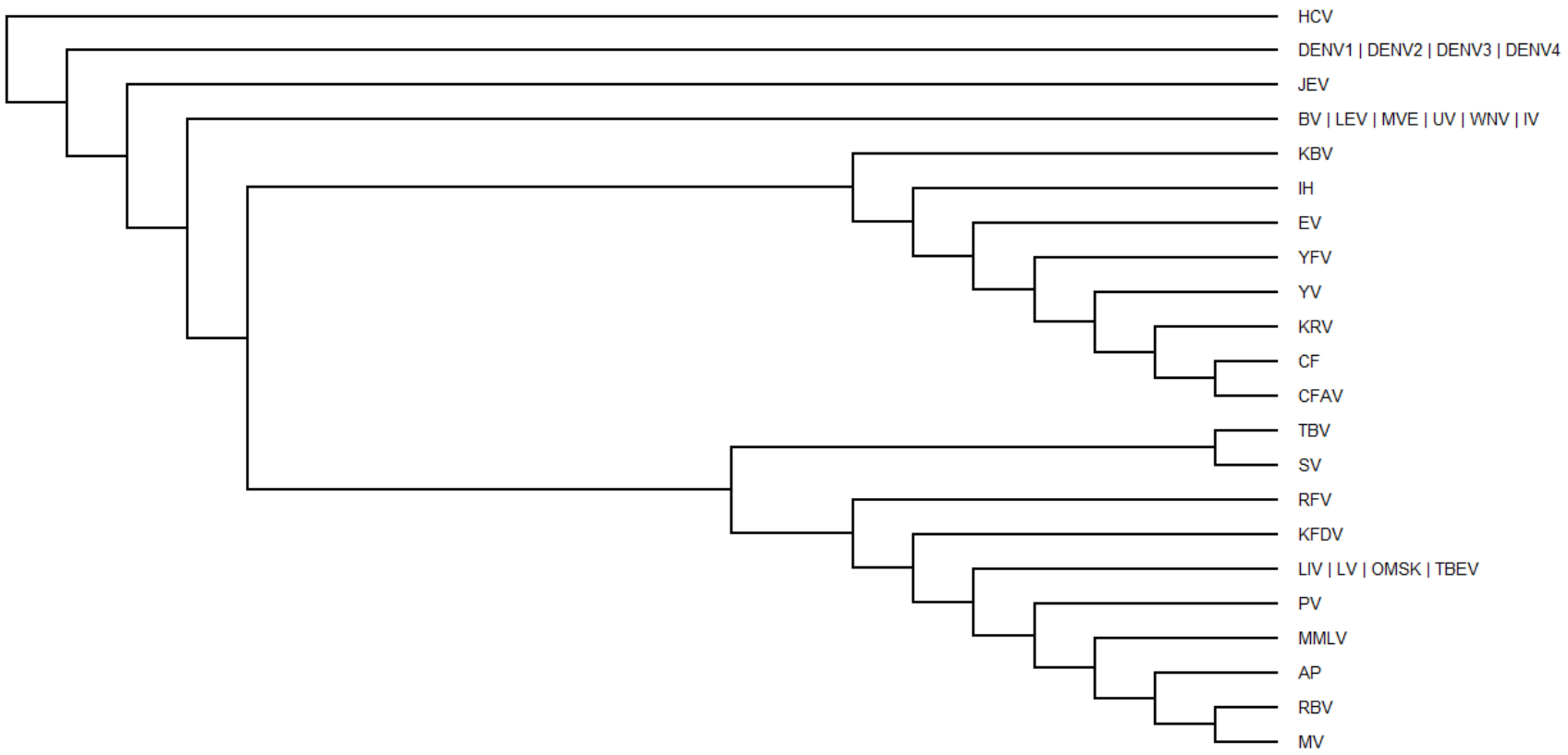
B) Evolution of E protein of selected *Flaviviruses*



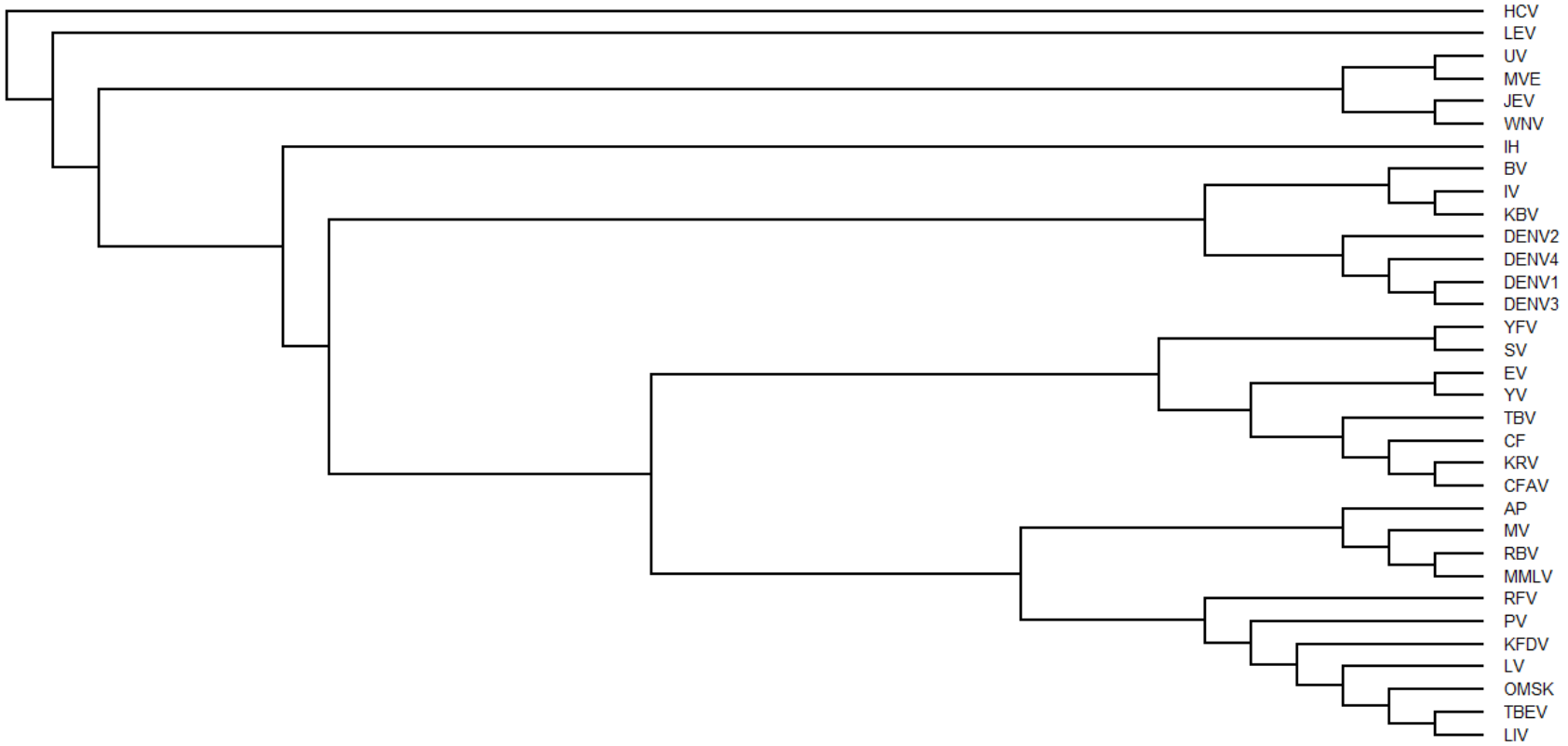
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence E $_{97}$VDRGWGNGCGLFGKG$_{111}$



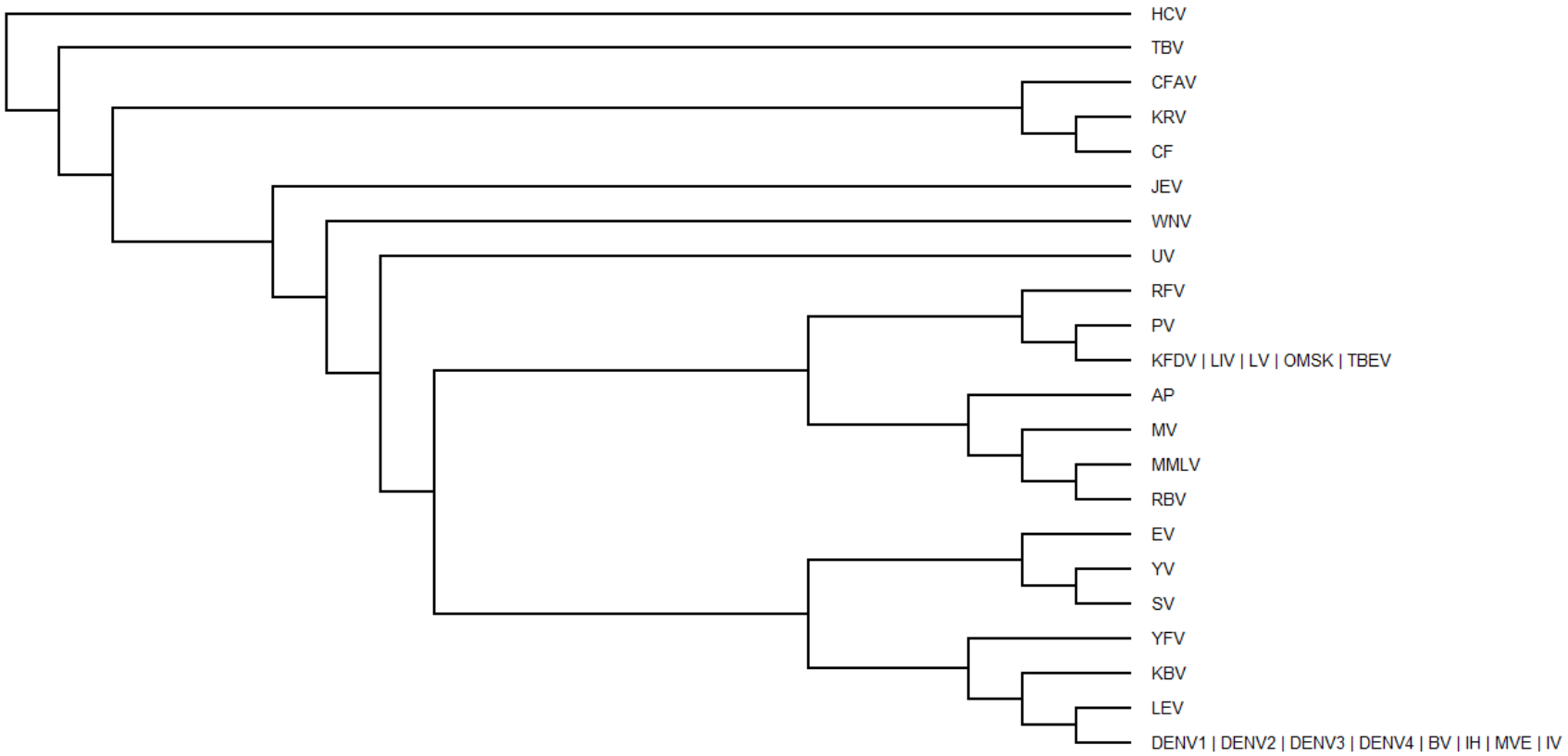
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence E $_{252}$VLGSQEGAMH$_{261}$



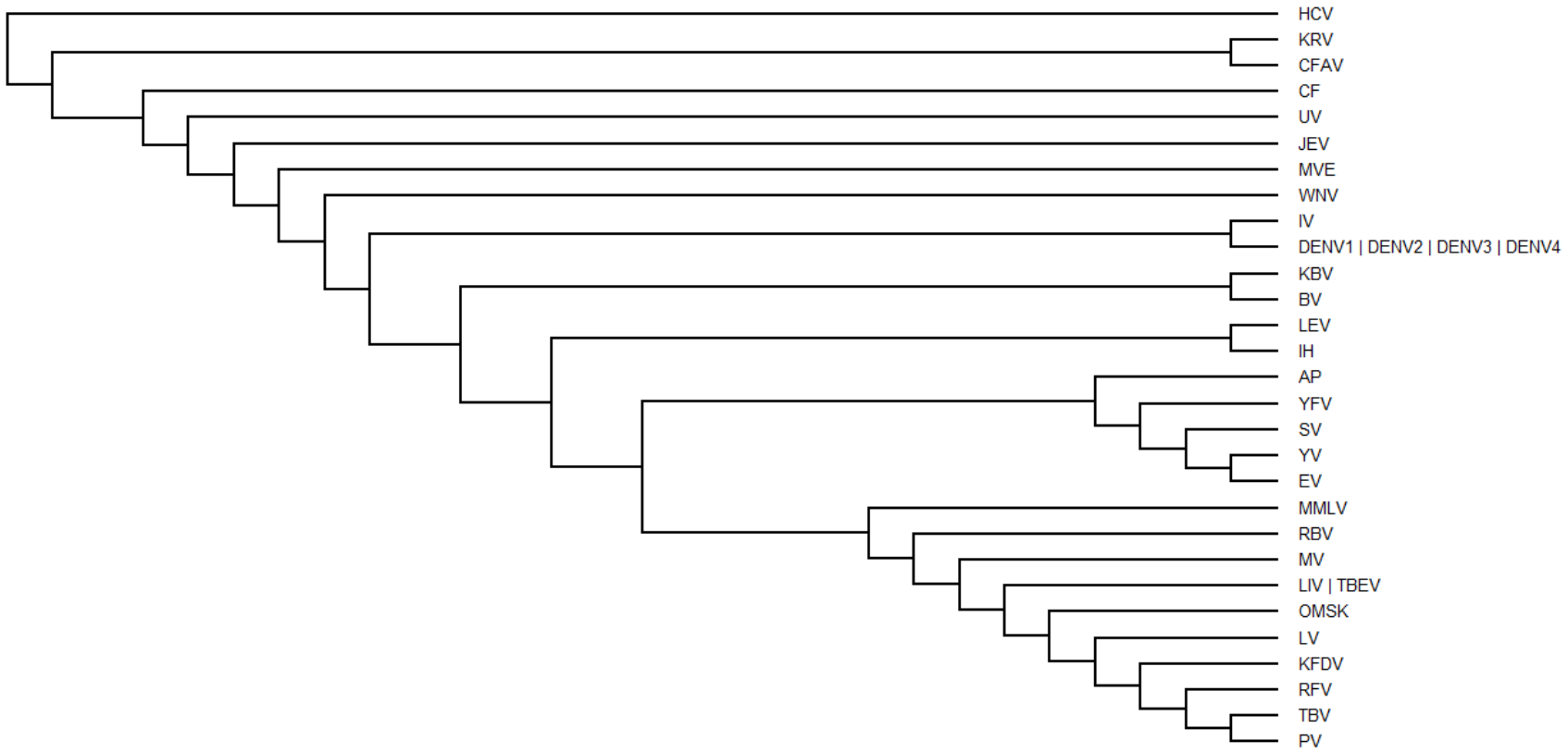
Evolution of NS1 protein of selected *Flaviviruses*



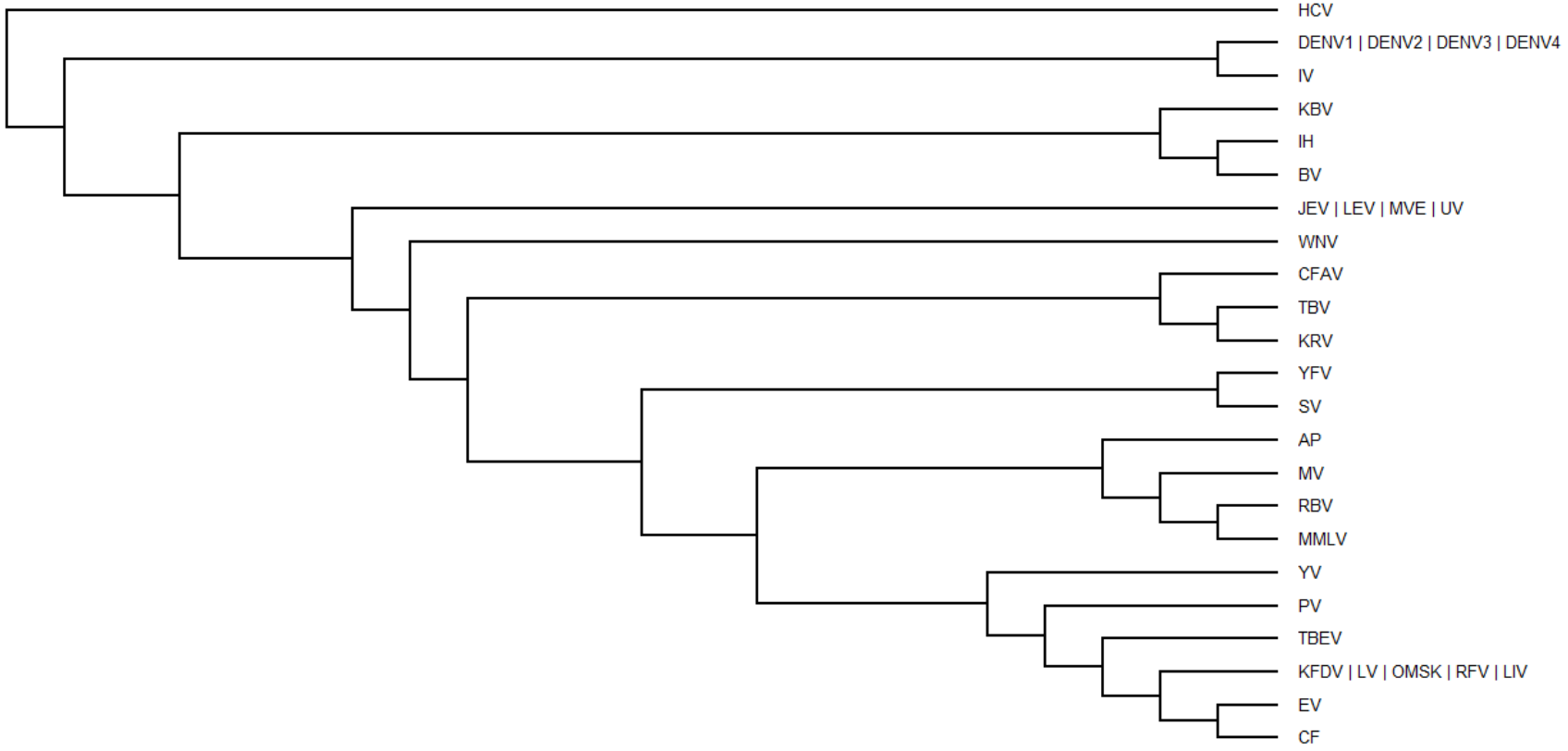
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{12}$ELKCGSGIF$_{20}$



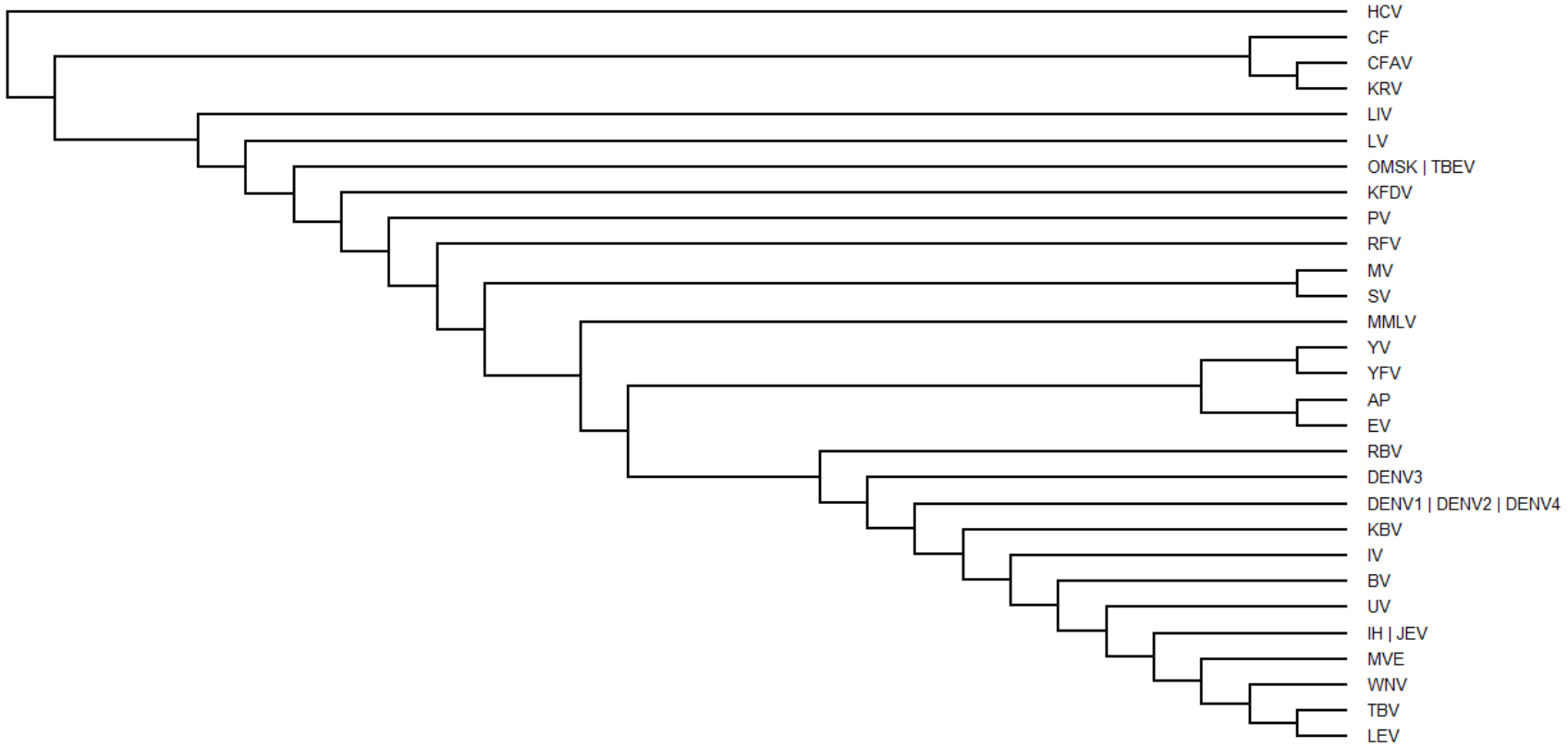
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{25}$VHTWTEQYKFQ$_{35}$



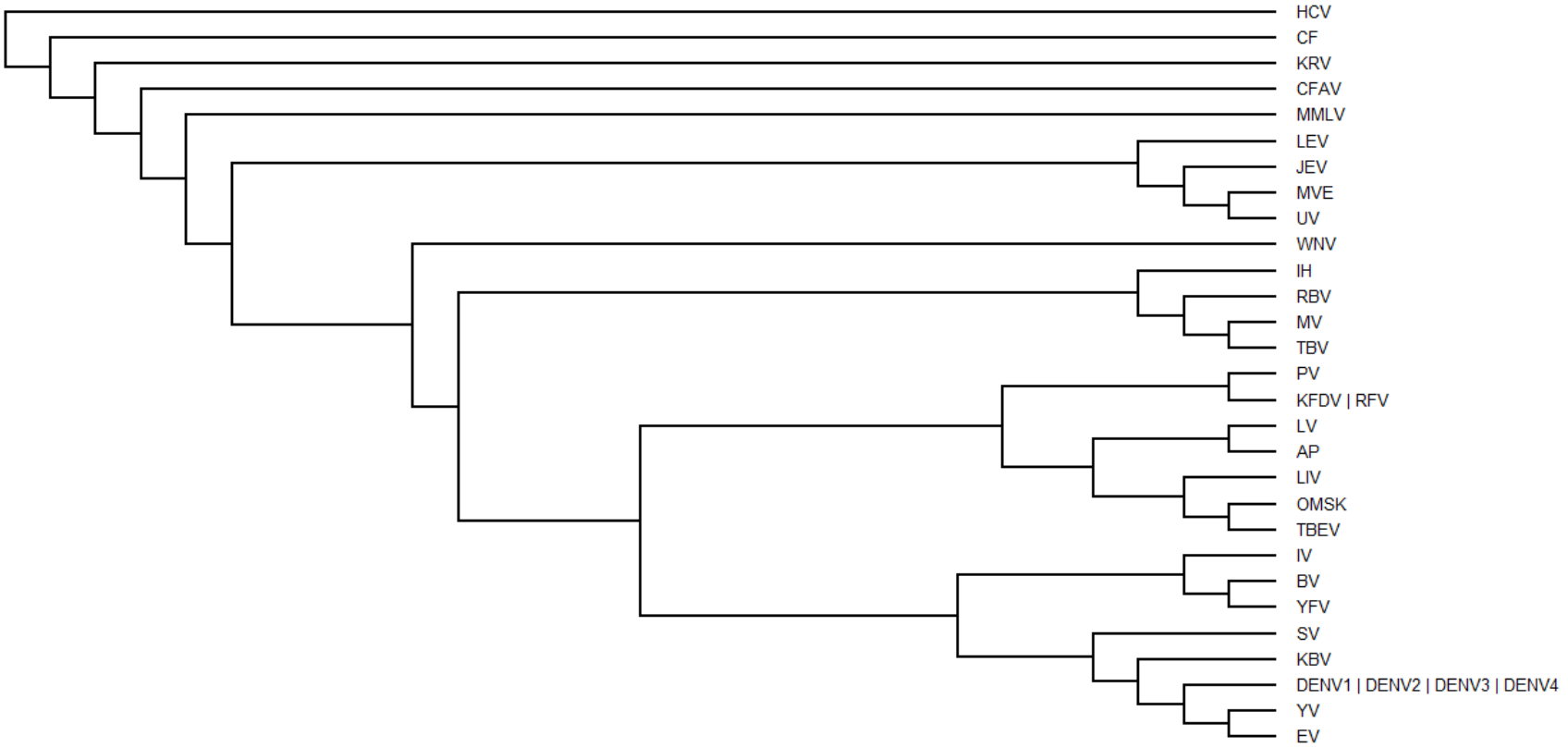
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{193}$AVHADMGYWIES$_{204}$



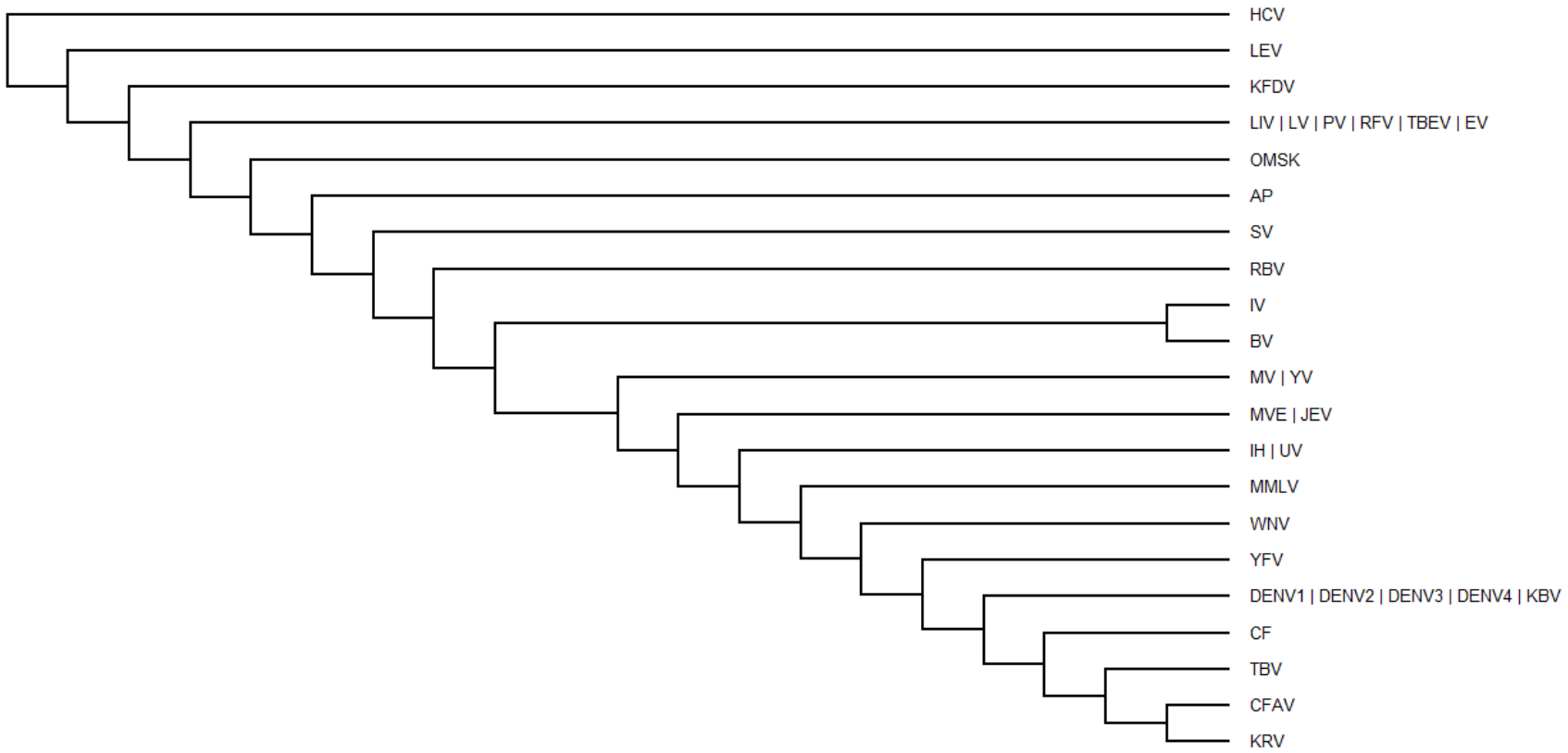
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{229}$HTLWSNGVLES$_{239}$



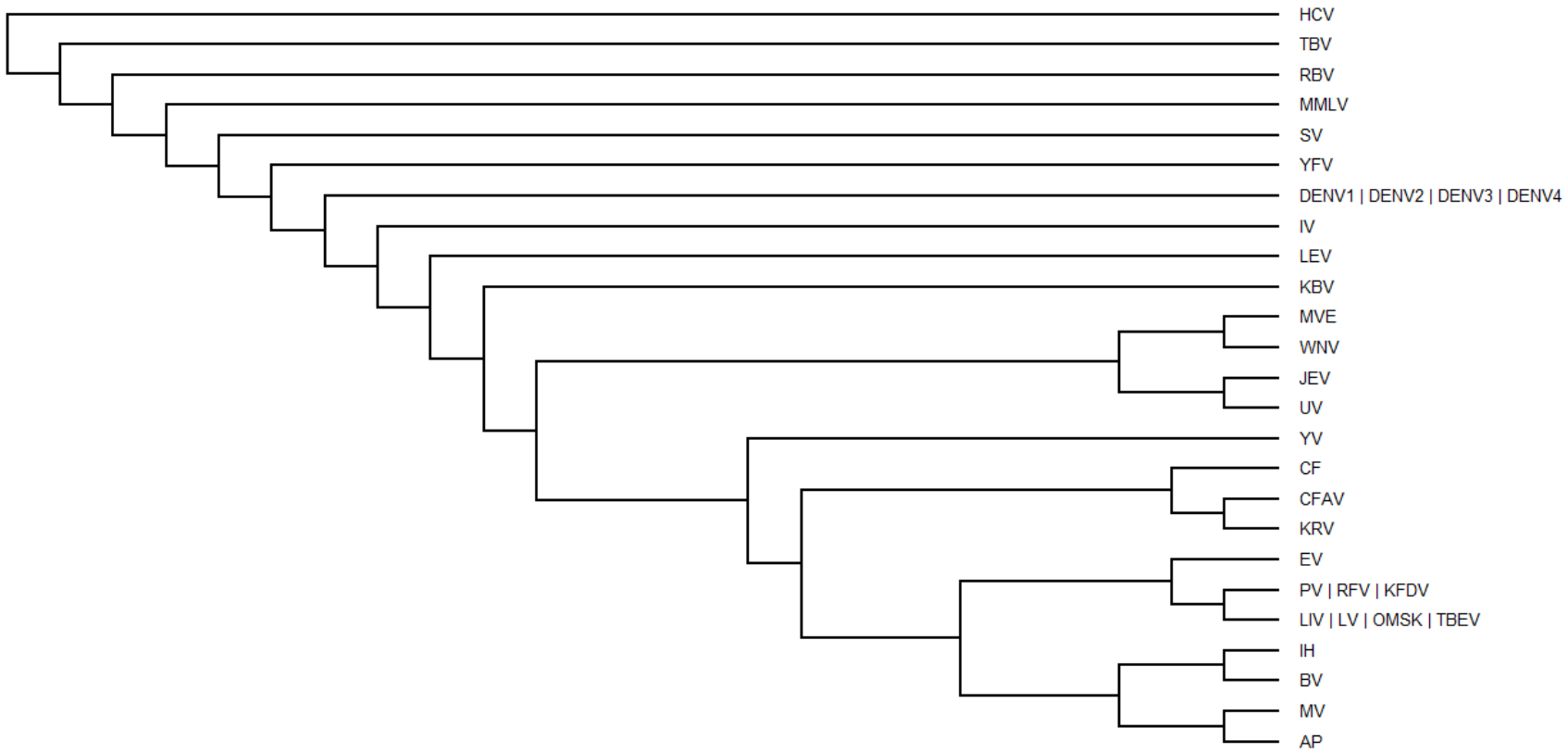
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{266}$GPWHLGKLE$_{274}$



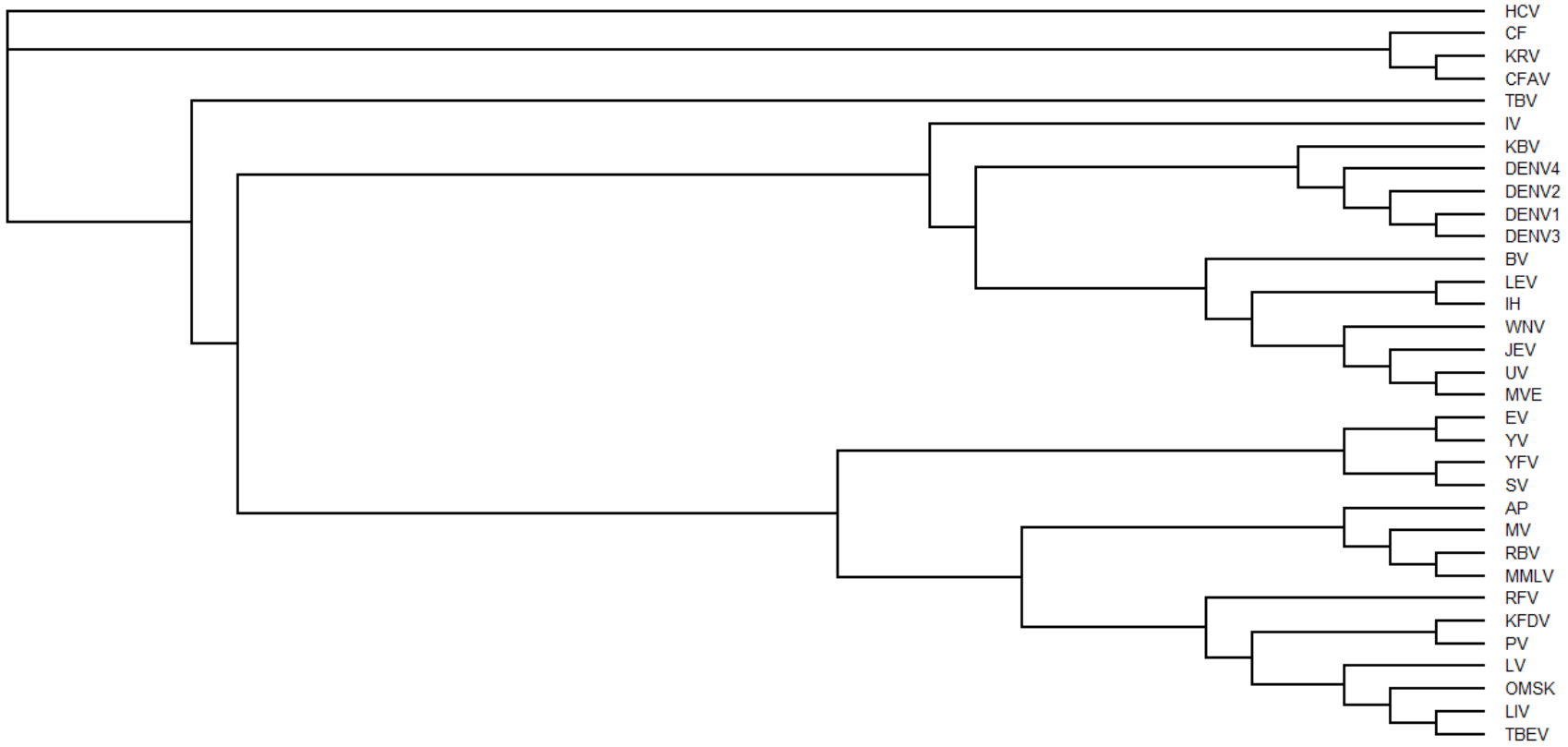
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{294}$RGPSLRTTT$_{302}$



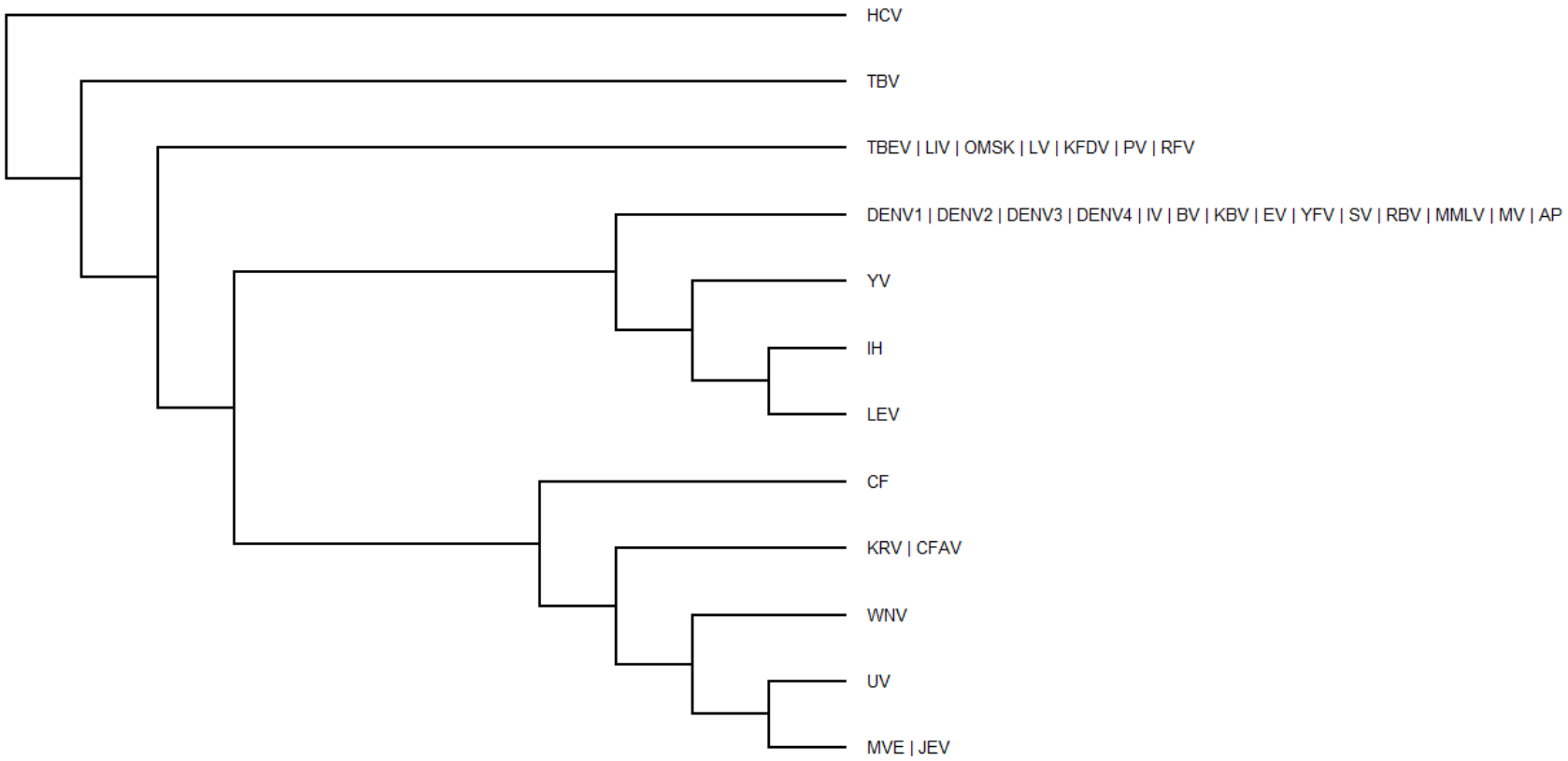
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS1 $_{325}$GEDGCWYGMEIRP$_{337}$



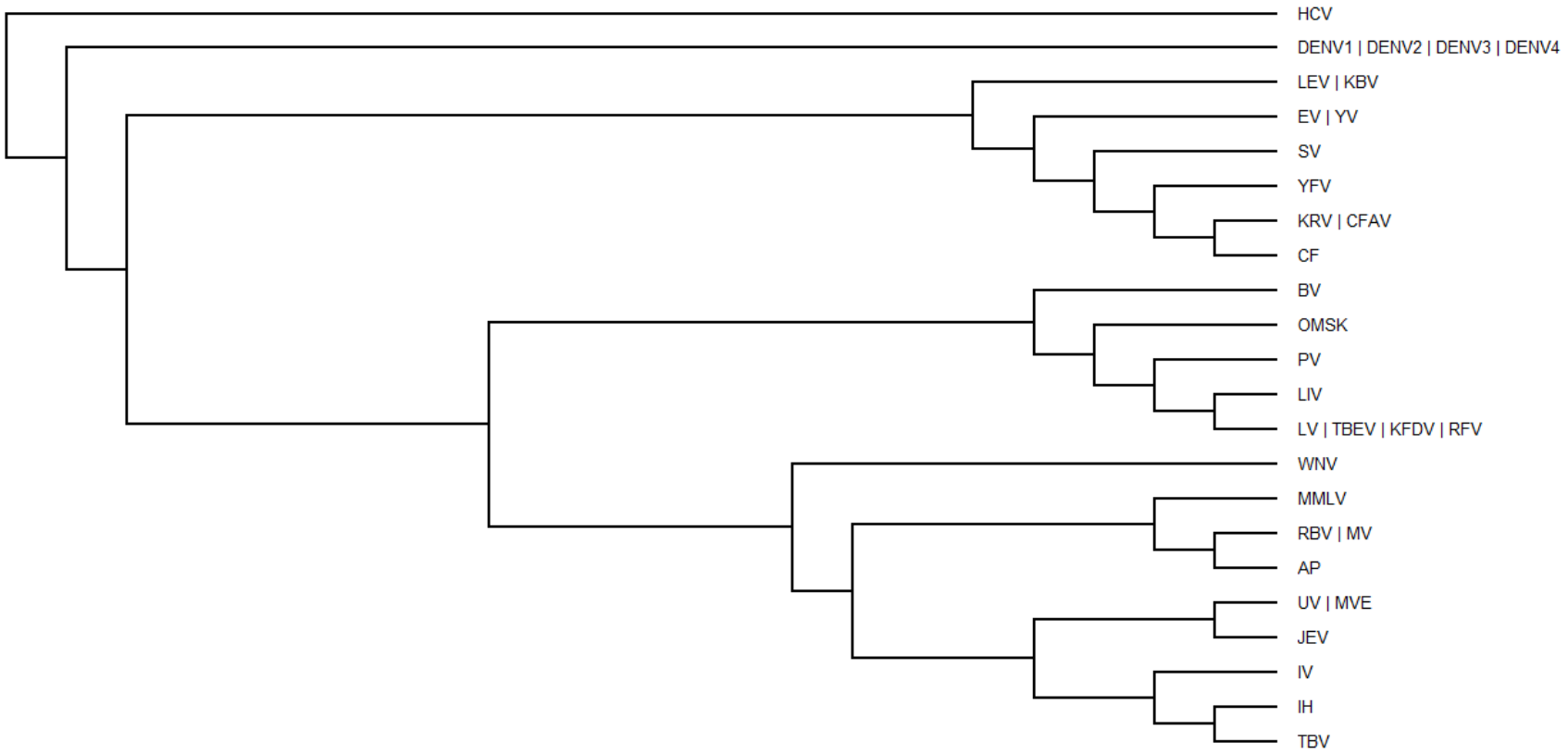
Evolution of NS3 protein of selected *Flaviviruses*



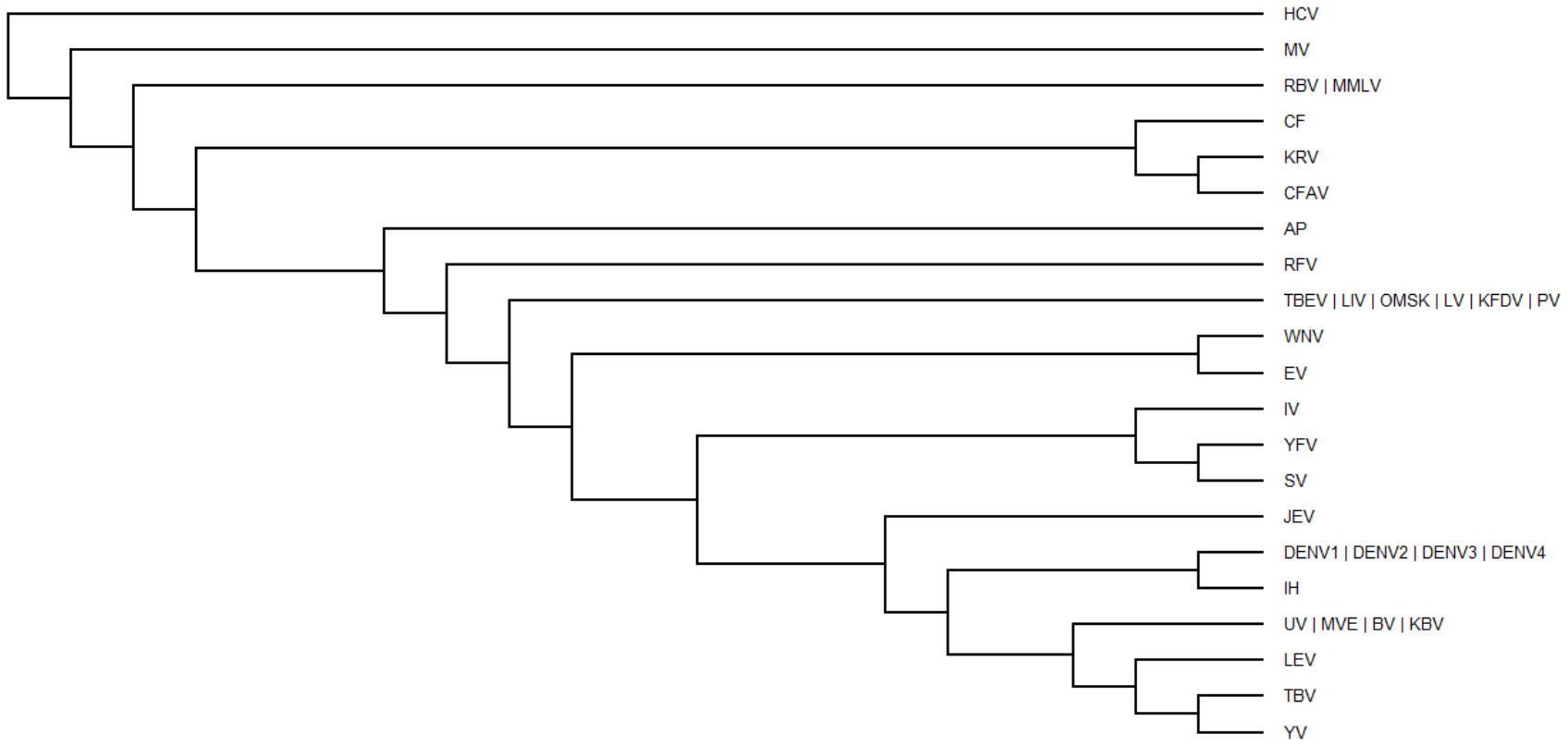
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{46}$FHTMWHVTRG$_{55}$



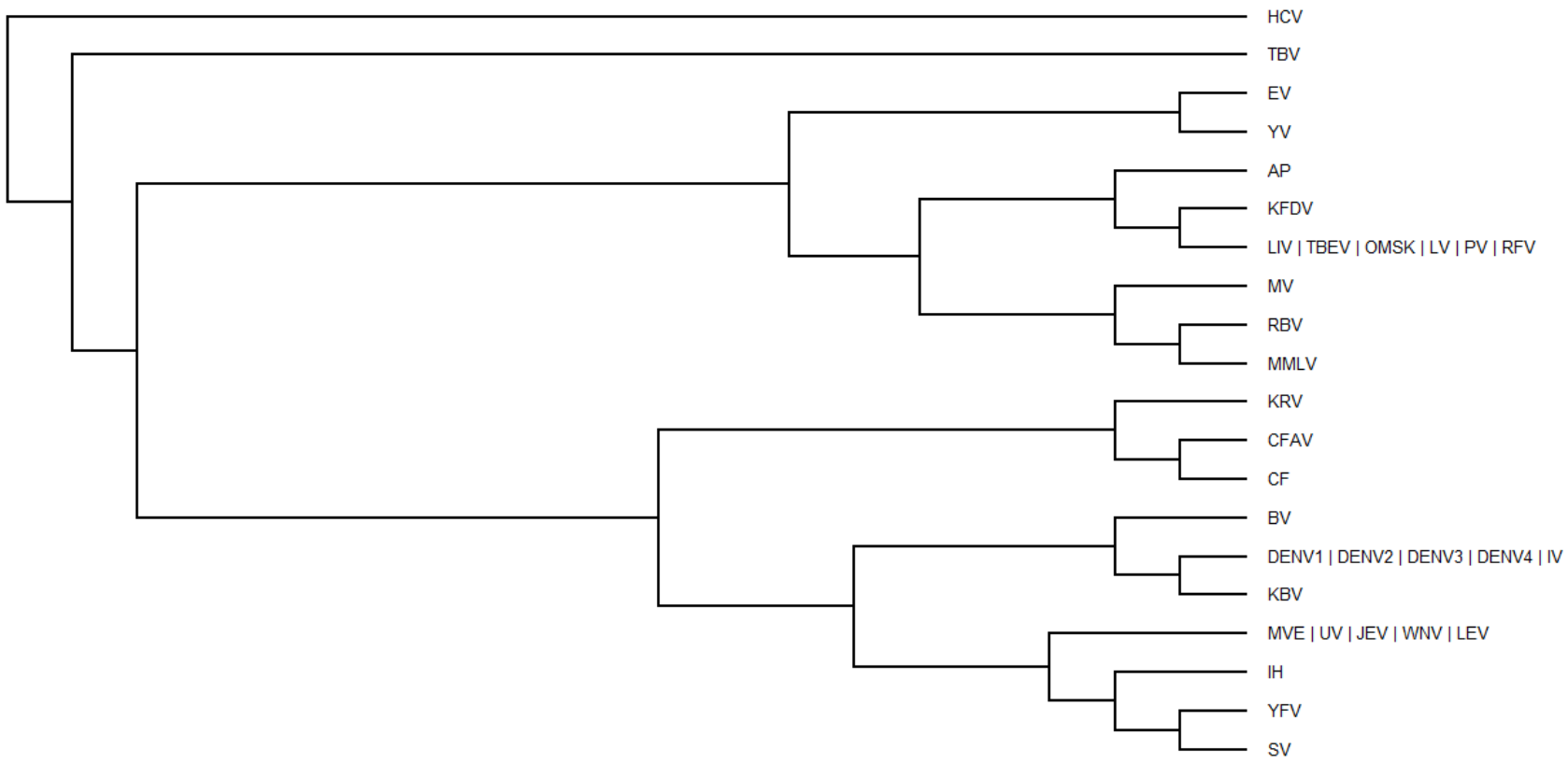
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{148}$GLYGNGVVT$_{156}$



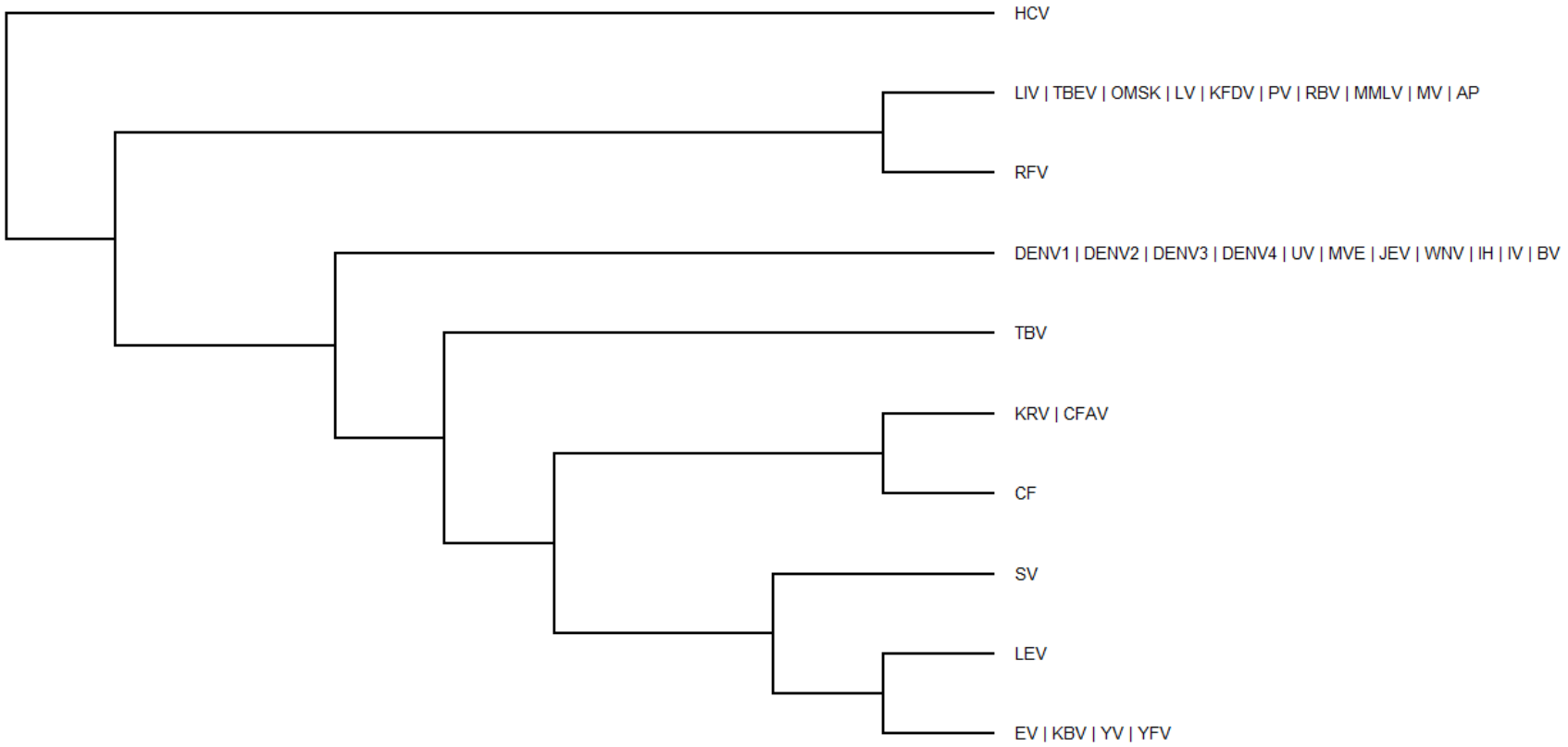
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{189}$LTIMDLHPG$_{197}$



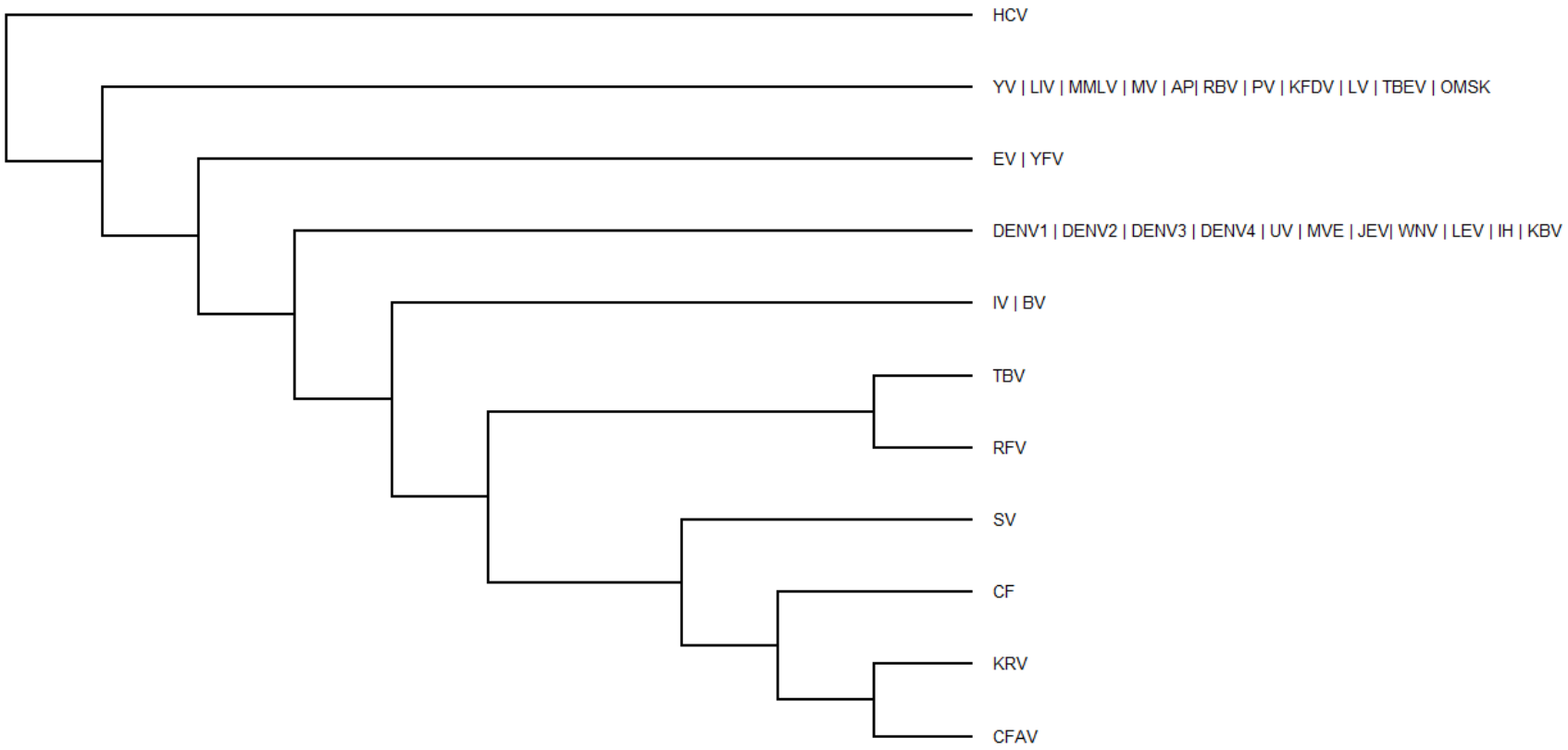
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{256}$EIVDLMCHATFT$_{267}$



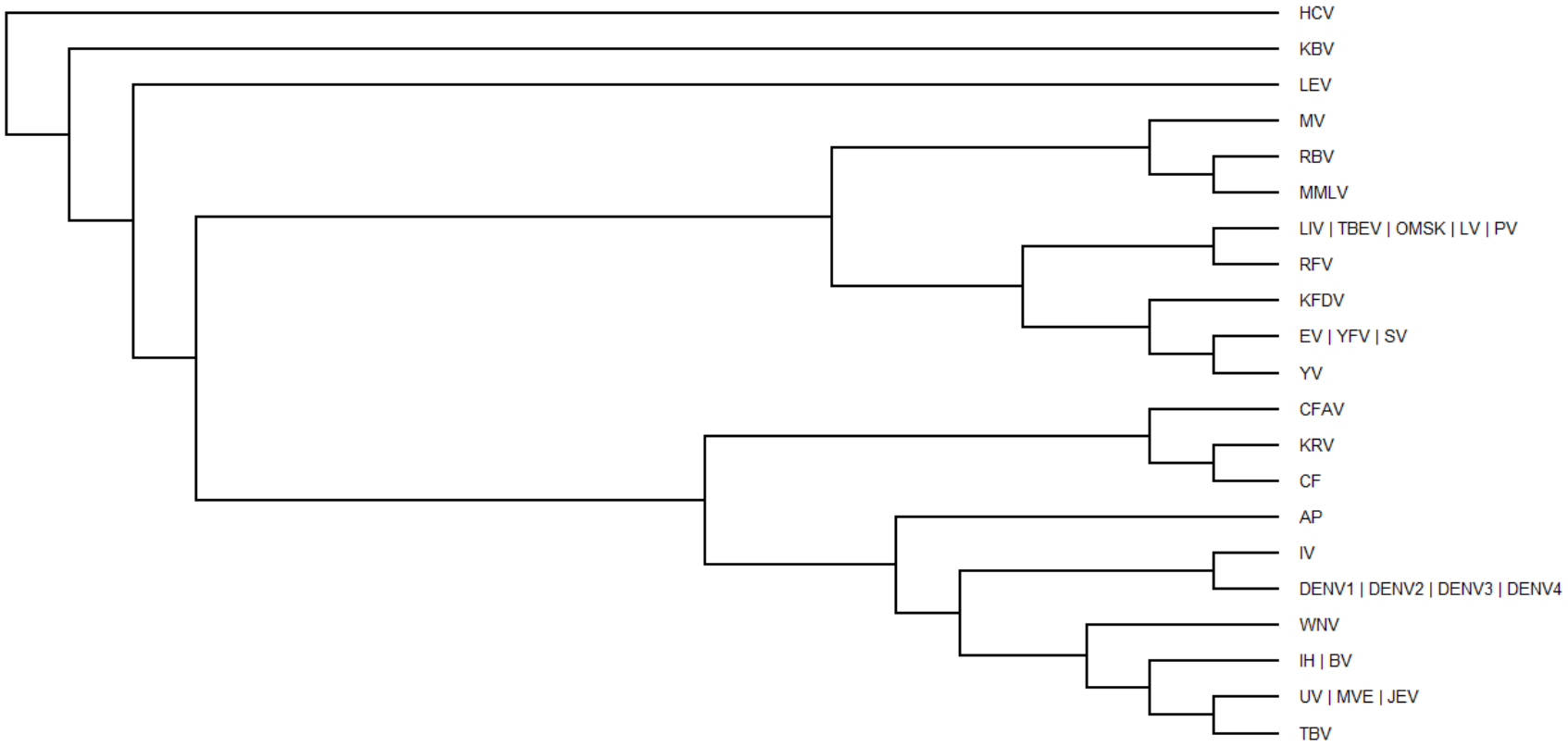
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{284}$MDEAHFTD$_{P292}$



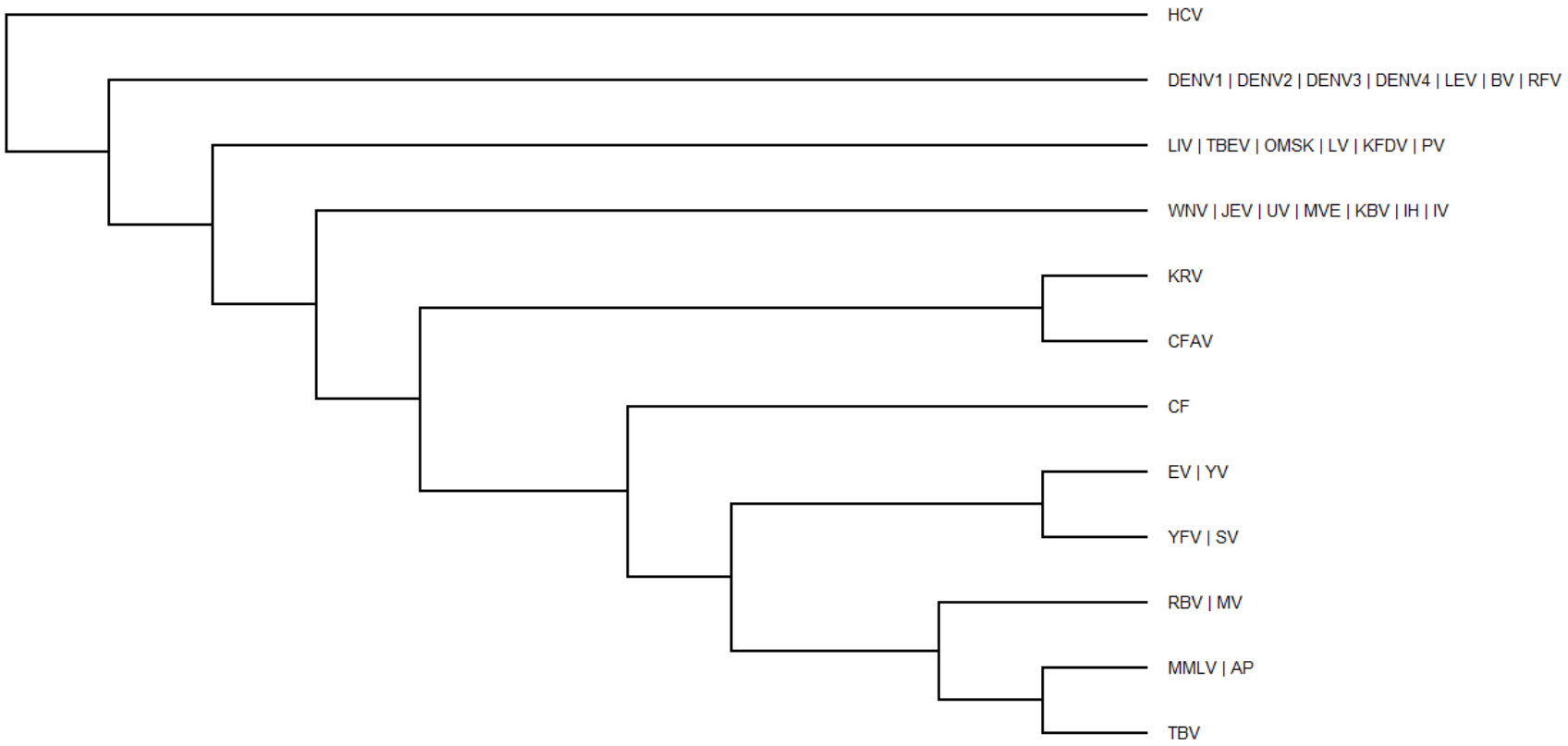
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{313}$IFMTATPPG$_{321}$



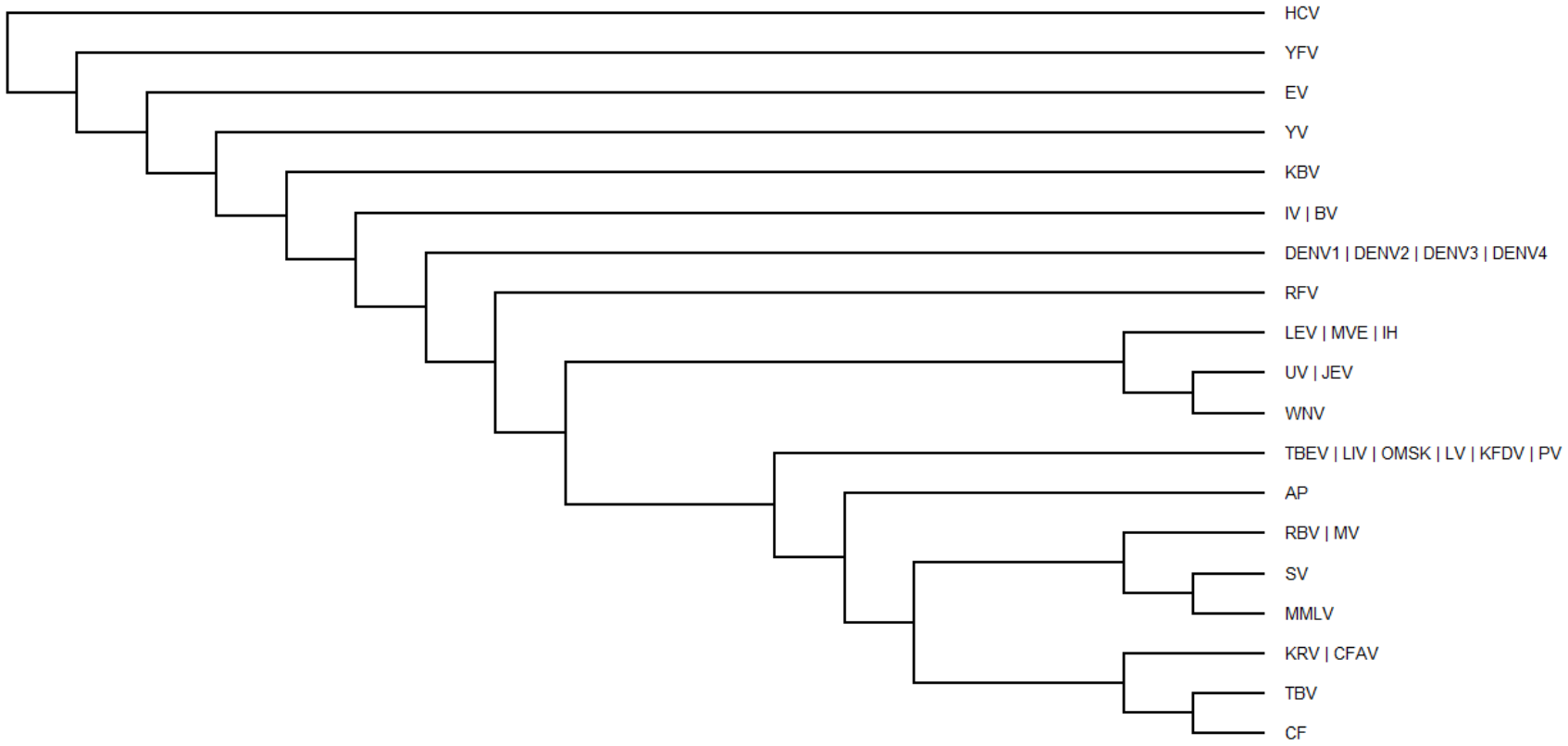
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{357}$GKTVWFVPSIK$_{367}$



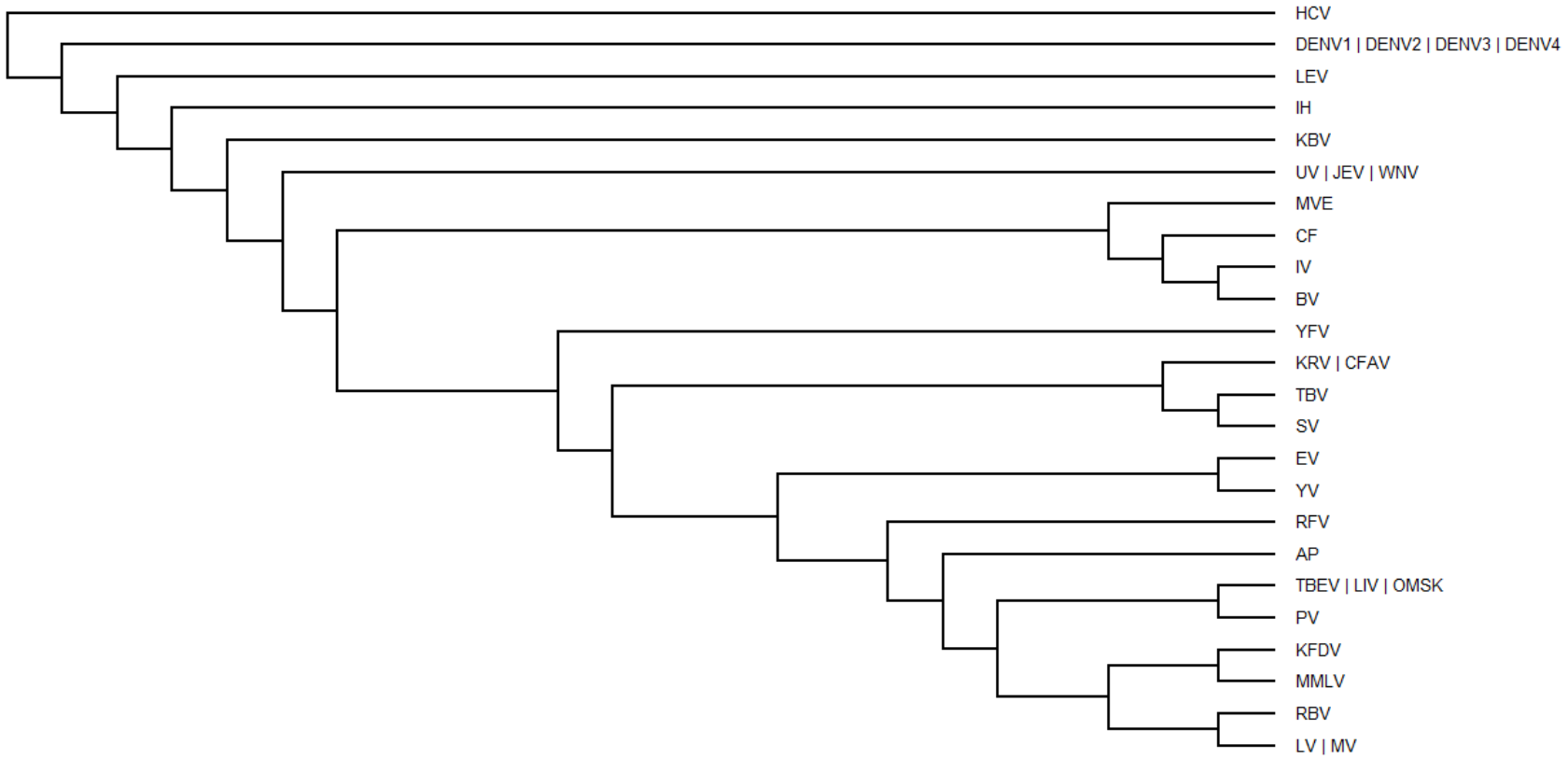
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{406}$VVTTDISEMGANF$_{418}$



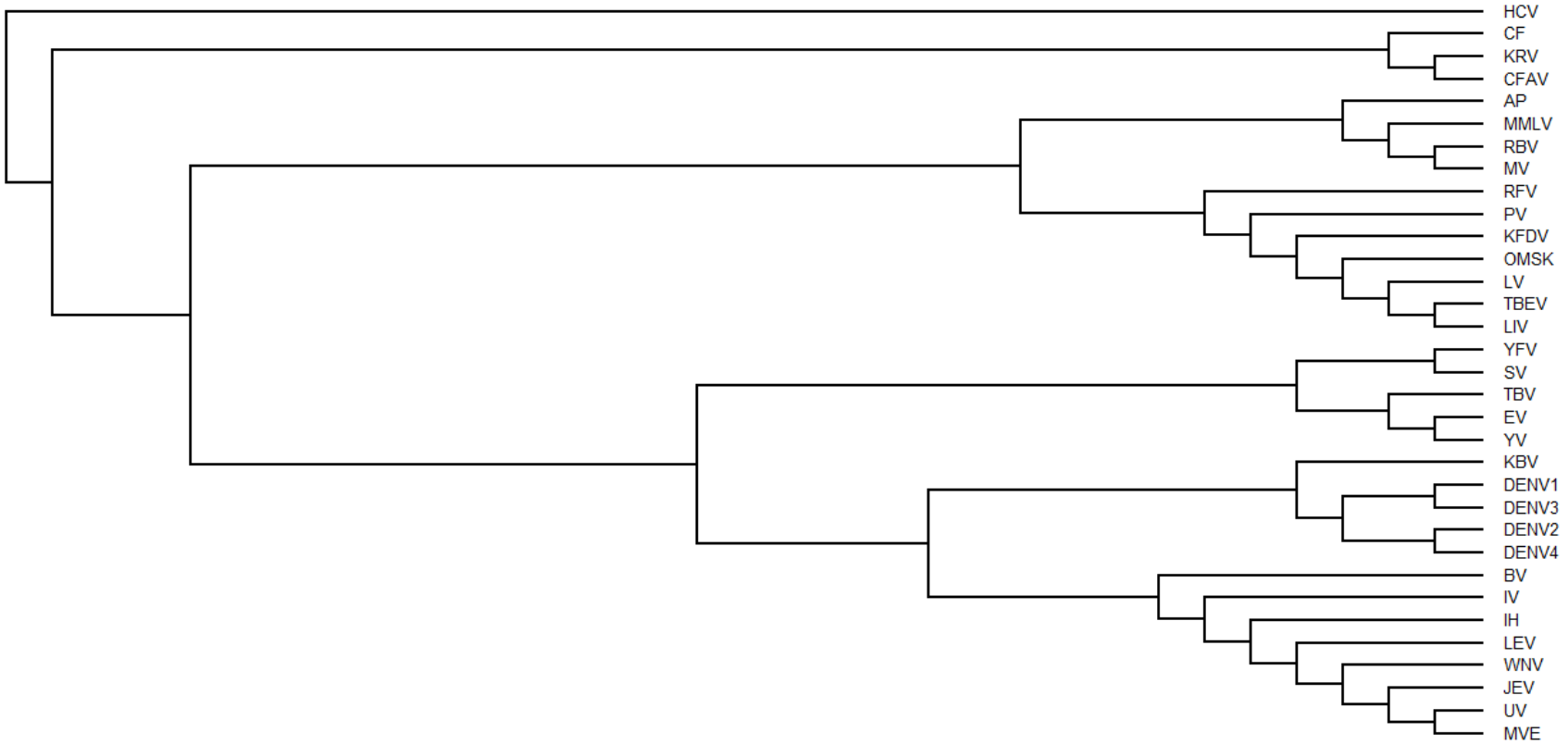
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{491}$EAKMLLDNI$_{499}$



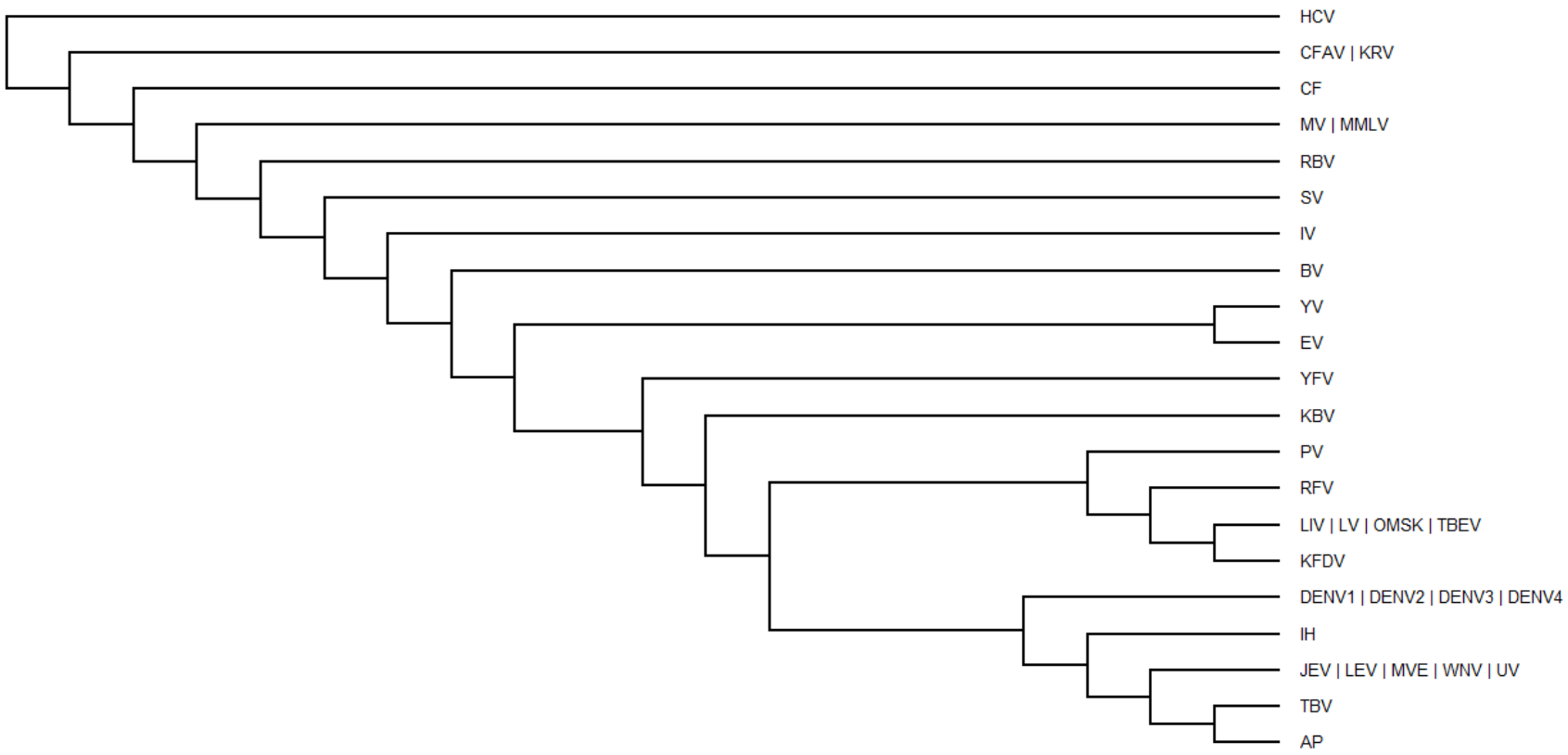
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS3 $_{537}$LMRRGDLPVWL$_{547}$



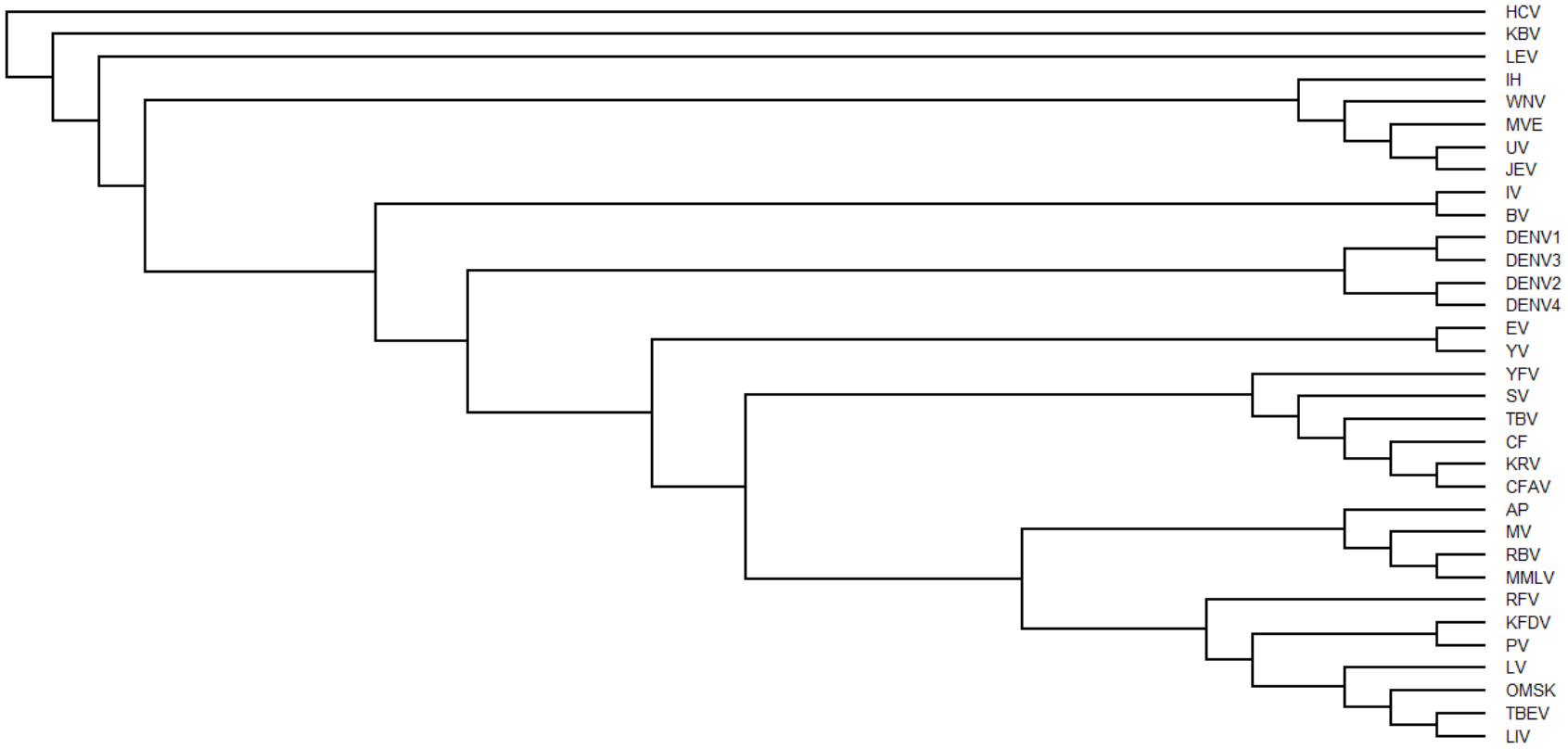
Evolution of NS4a protein of selected *Flaviviruses*



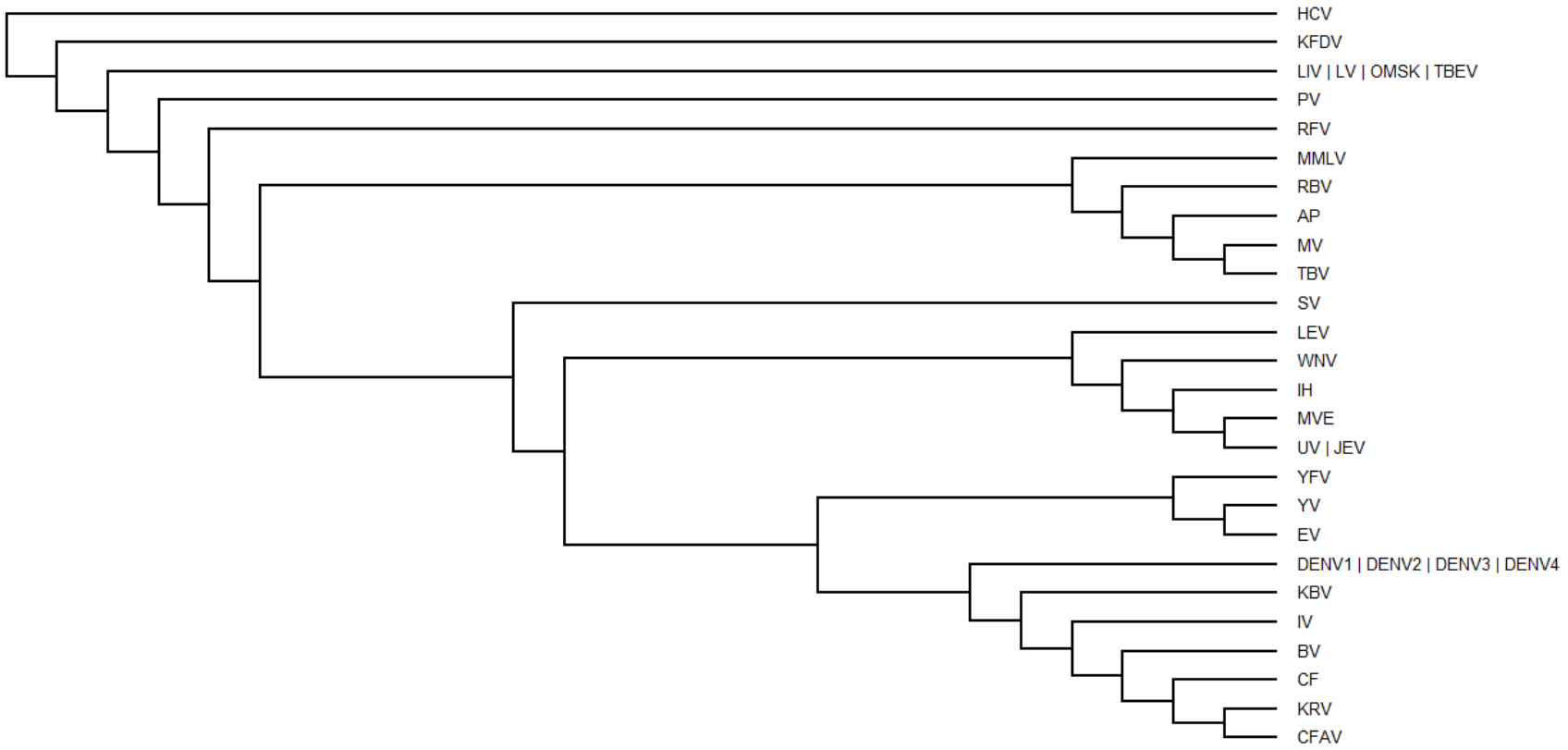
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS4a $_{126}$QRTPQDNQL$_{134}$



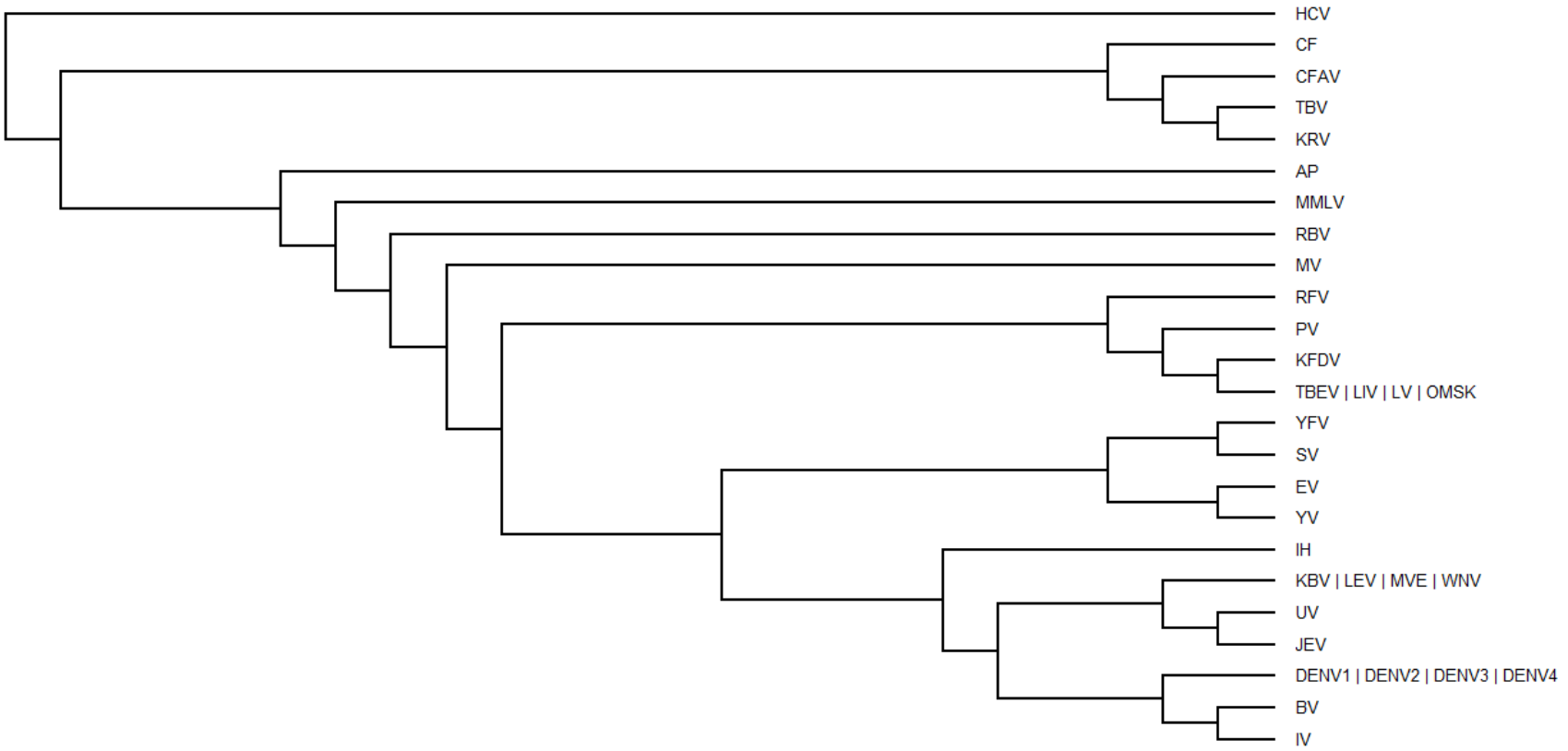
Evolution of NS4b protein of selected *Flaviviruses*



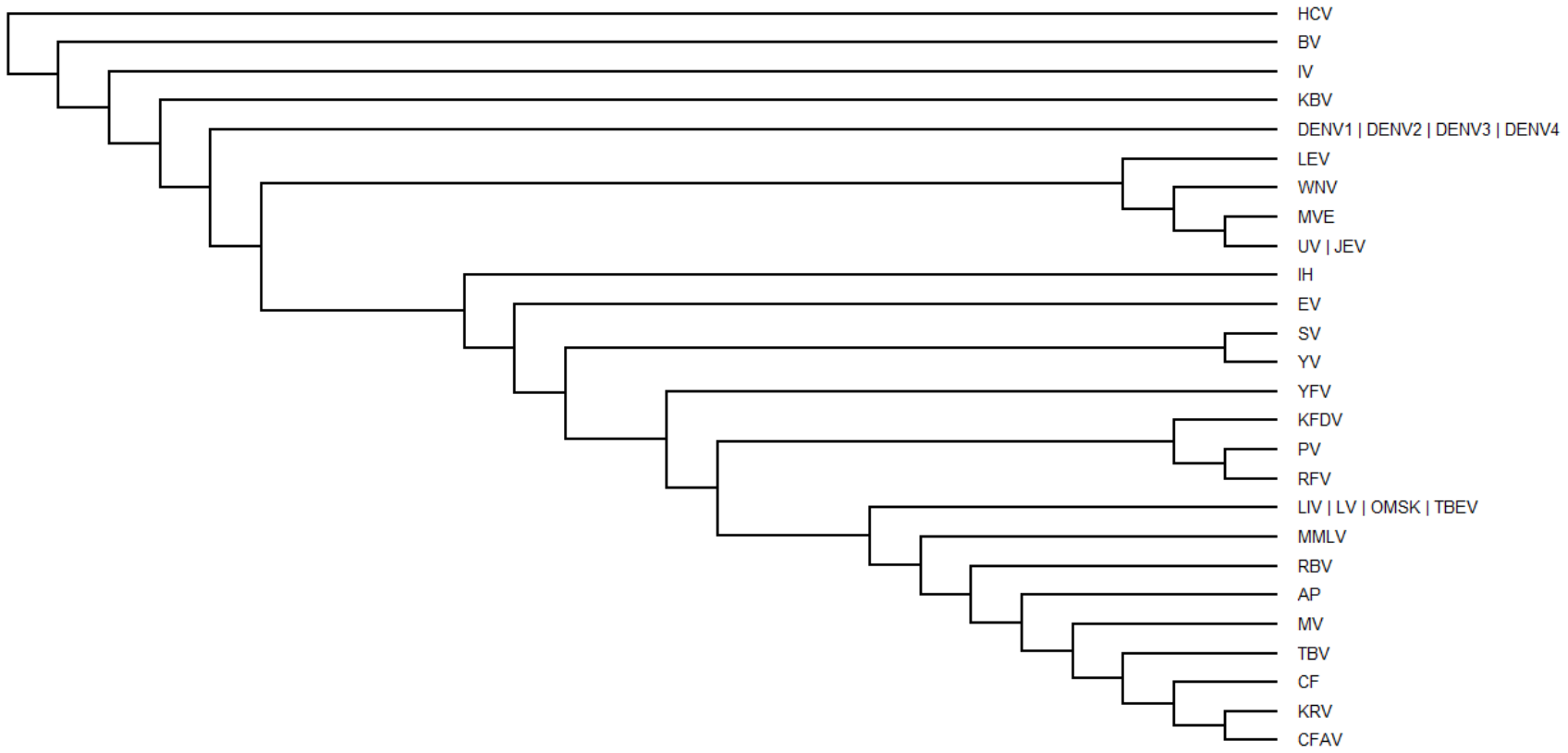
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS4b $_{118}$HYAIIGPGLQAKATREAQKR$_{137}$



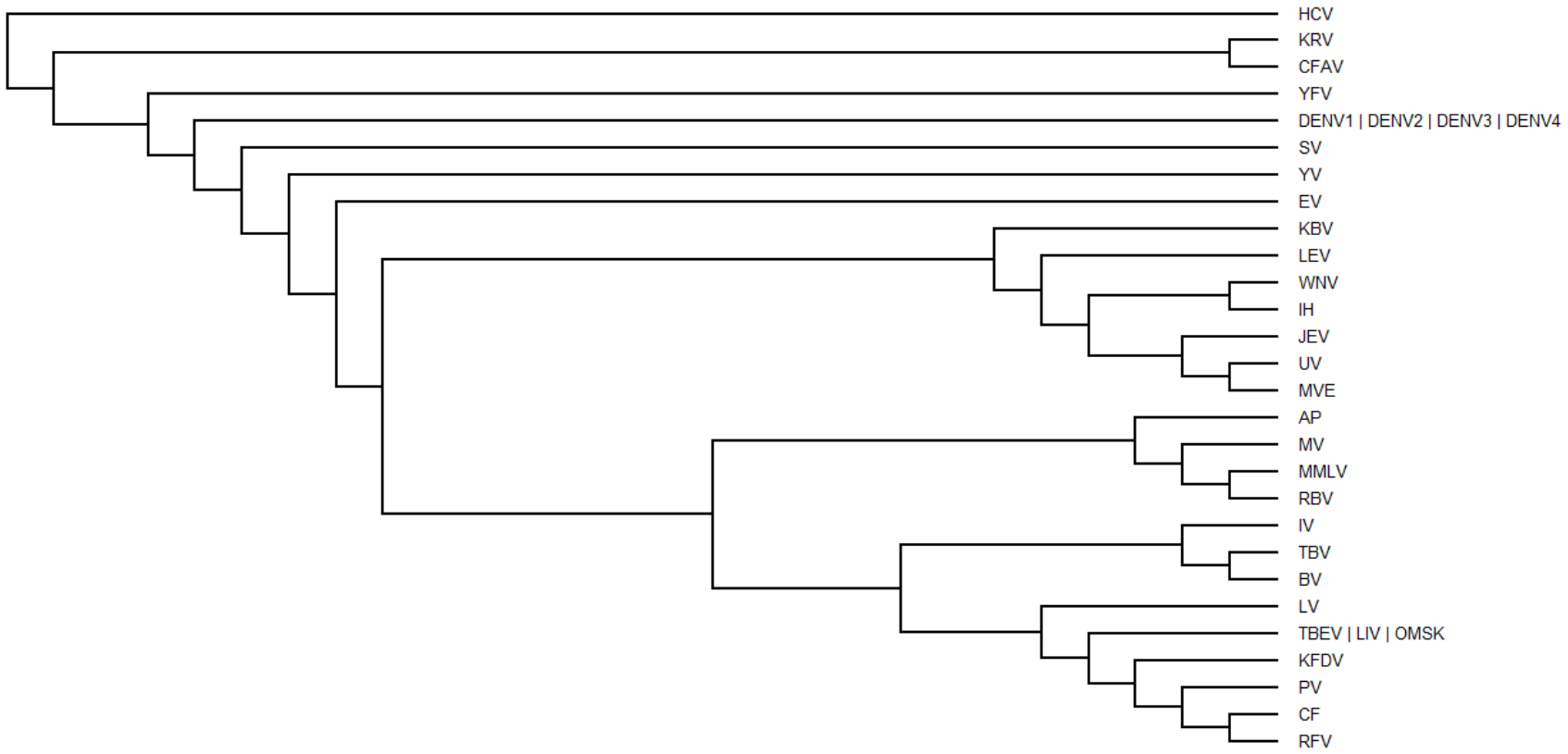
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS4b $_{139}$AAGIMKNPTVDGI$_{151}$



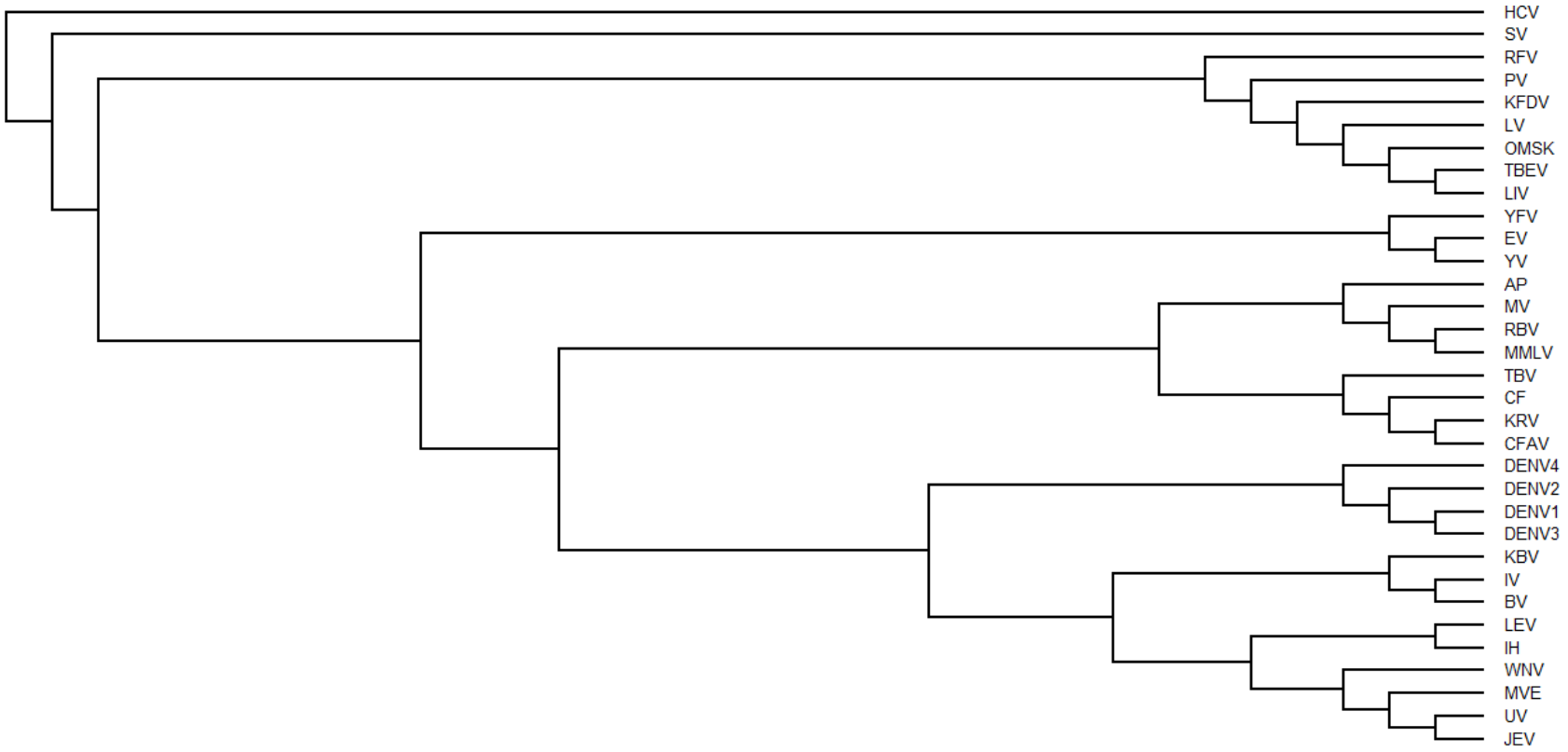
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS4b $_{213}$FWNTTIAVS$_{221}$



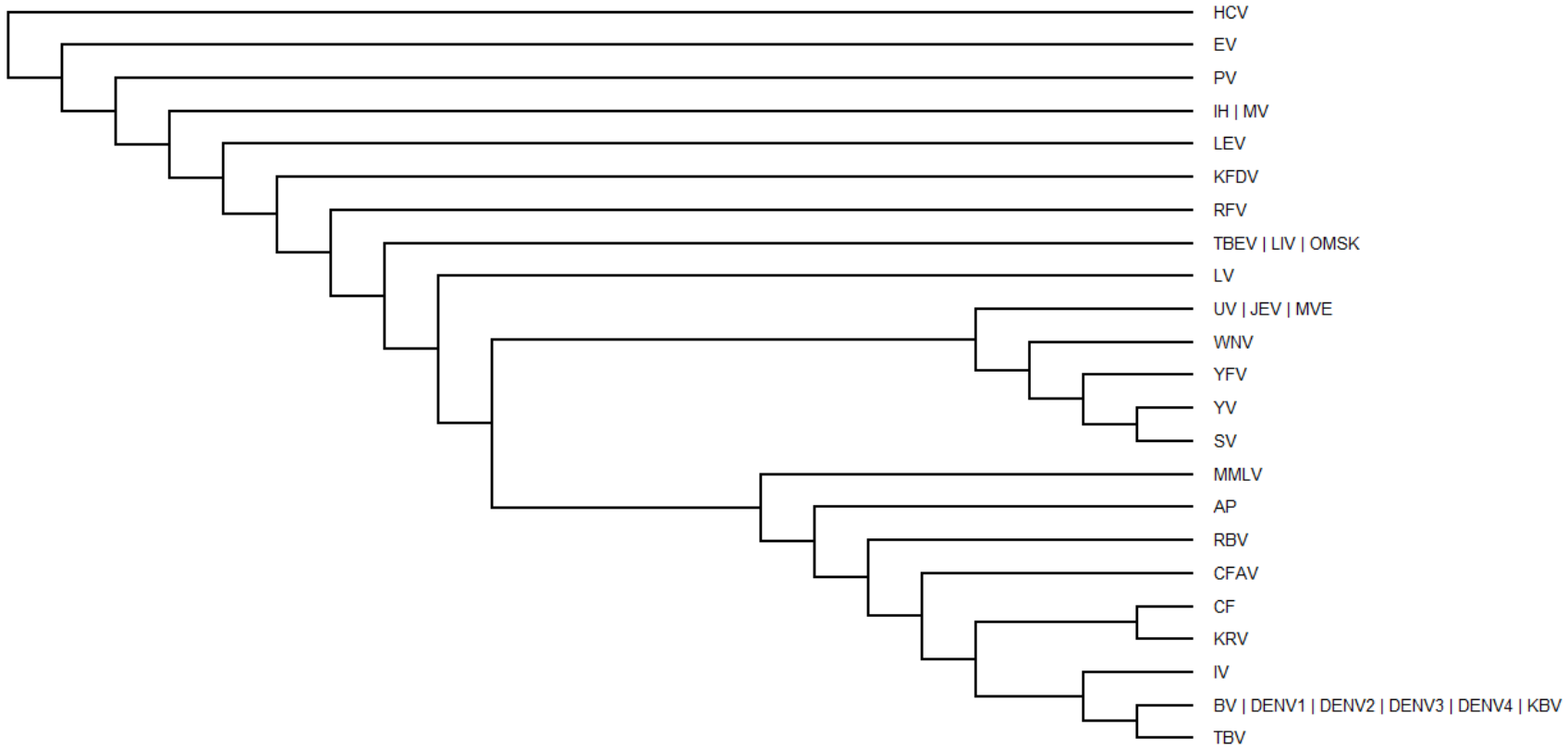
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS4b $_{223}$ANIFRGSYLAGAGL$_{236}$



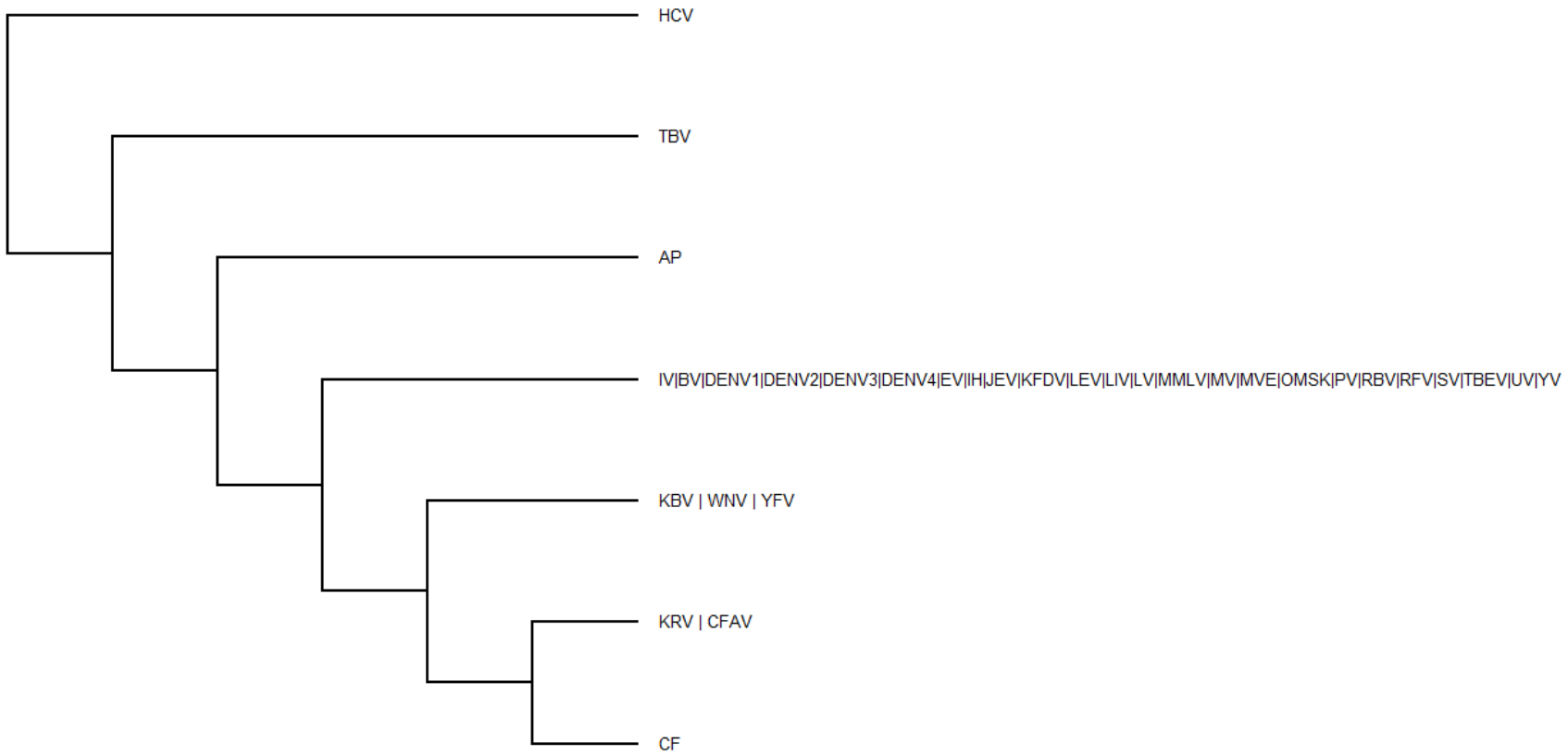
Evolution of NS5 protein of selected *Flaviviruses*



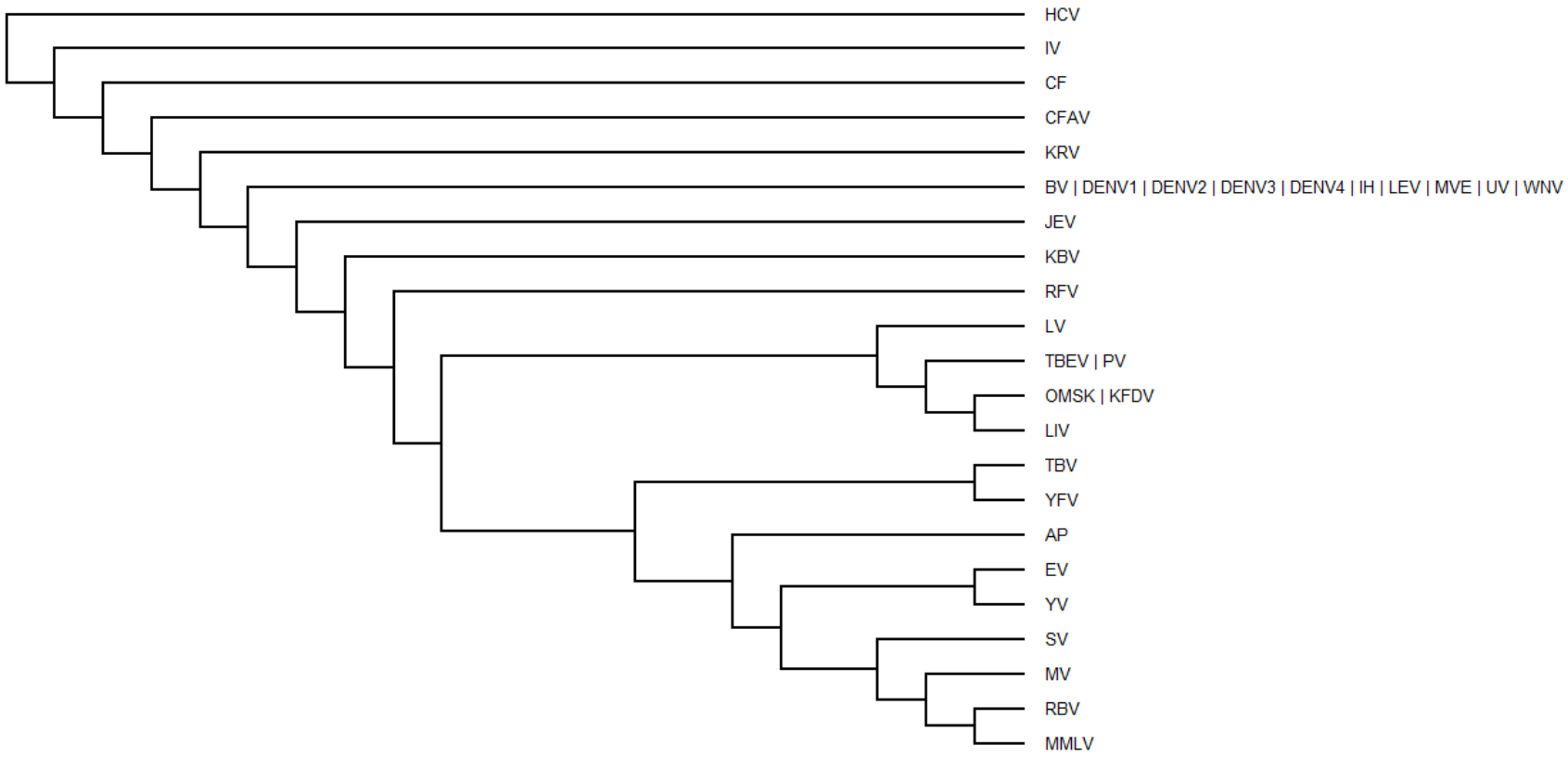
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{6}$GETLGEKWK$_{14}$



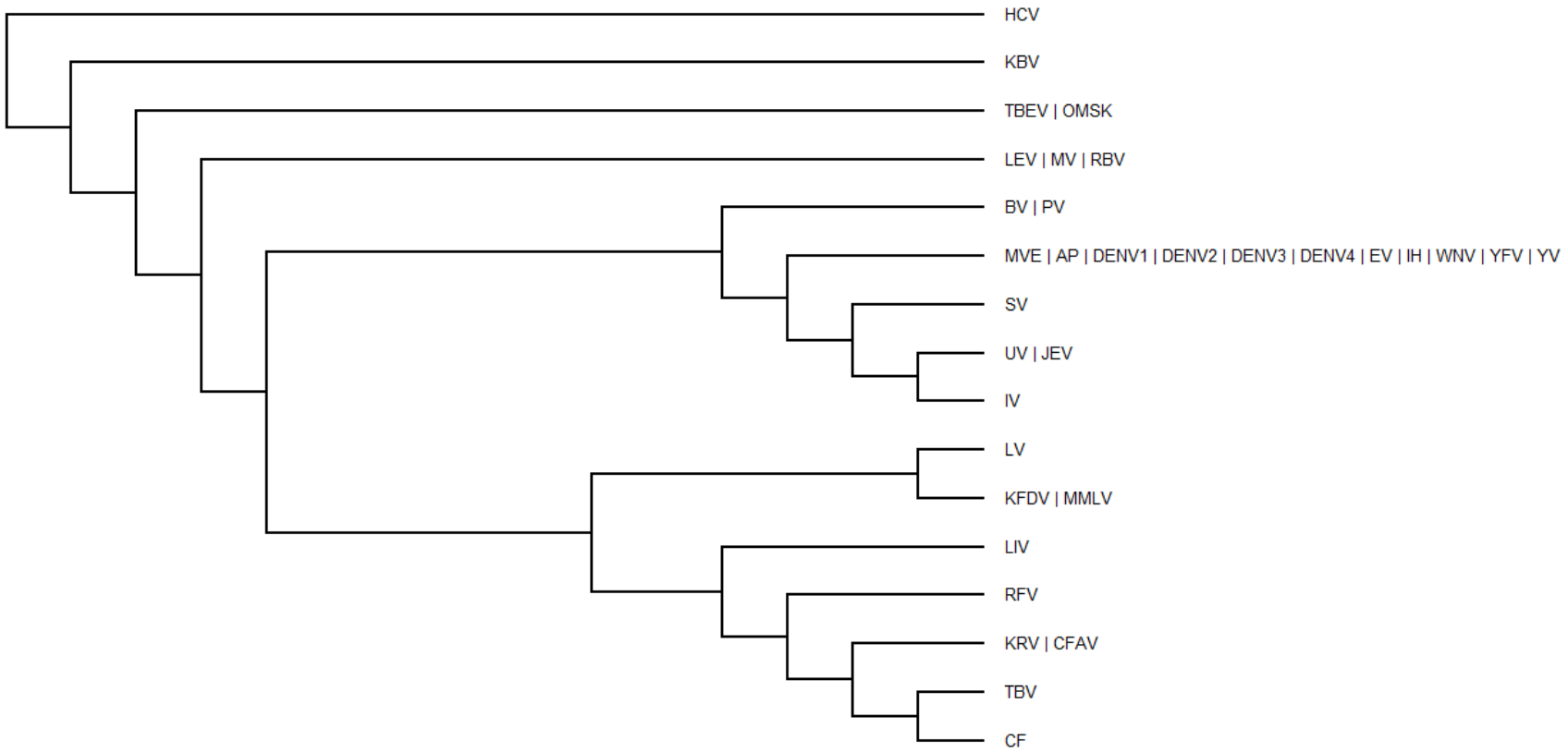
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{79}$DLGCGRGGWSYY$_{90}$



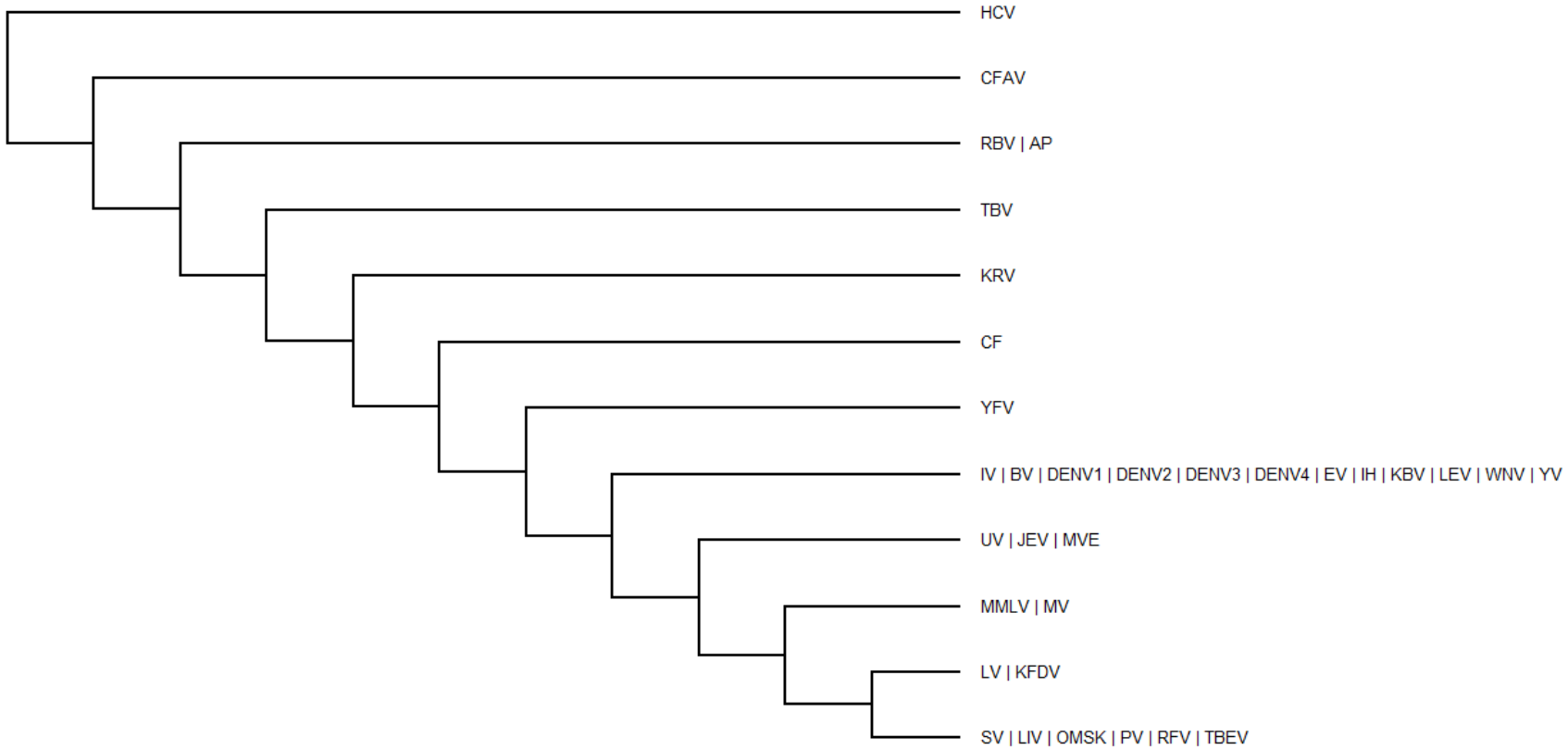
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{104}$TKGGPGHEEP$_{113}$



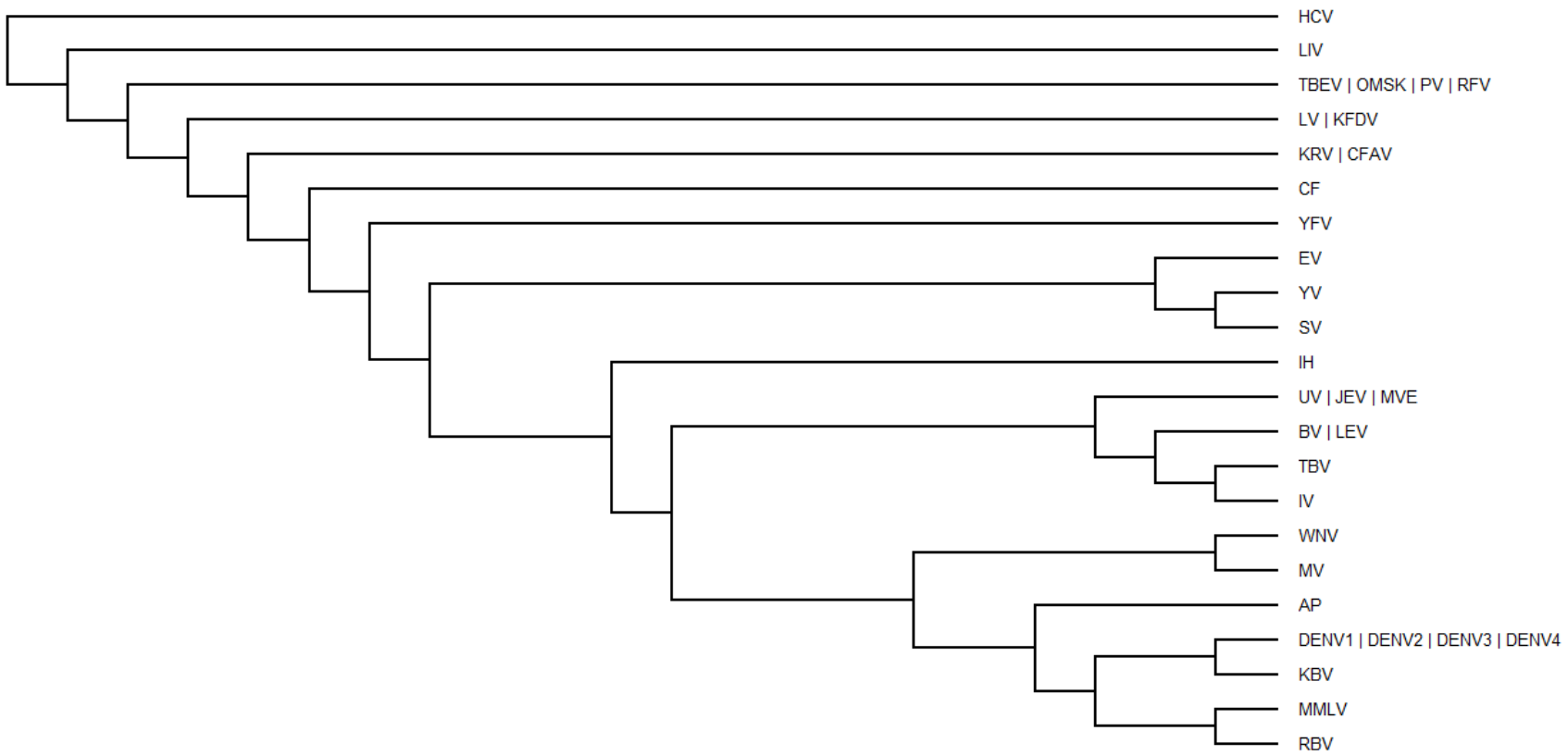
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{141}$DTLLCDIGESS$_{151}$



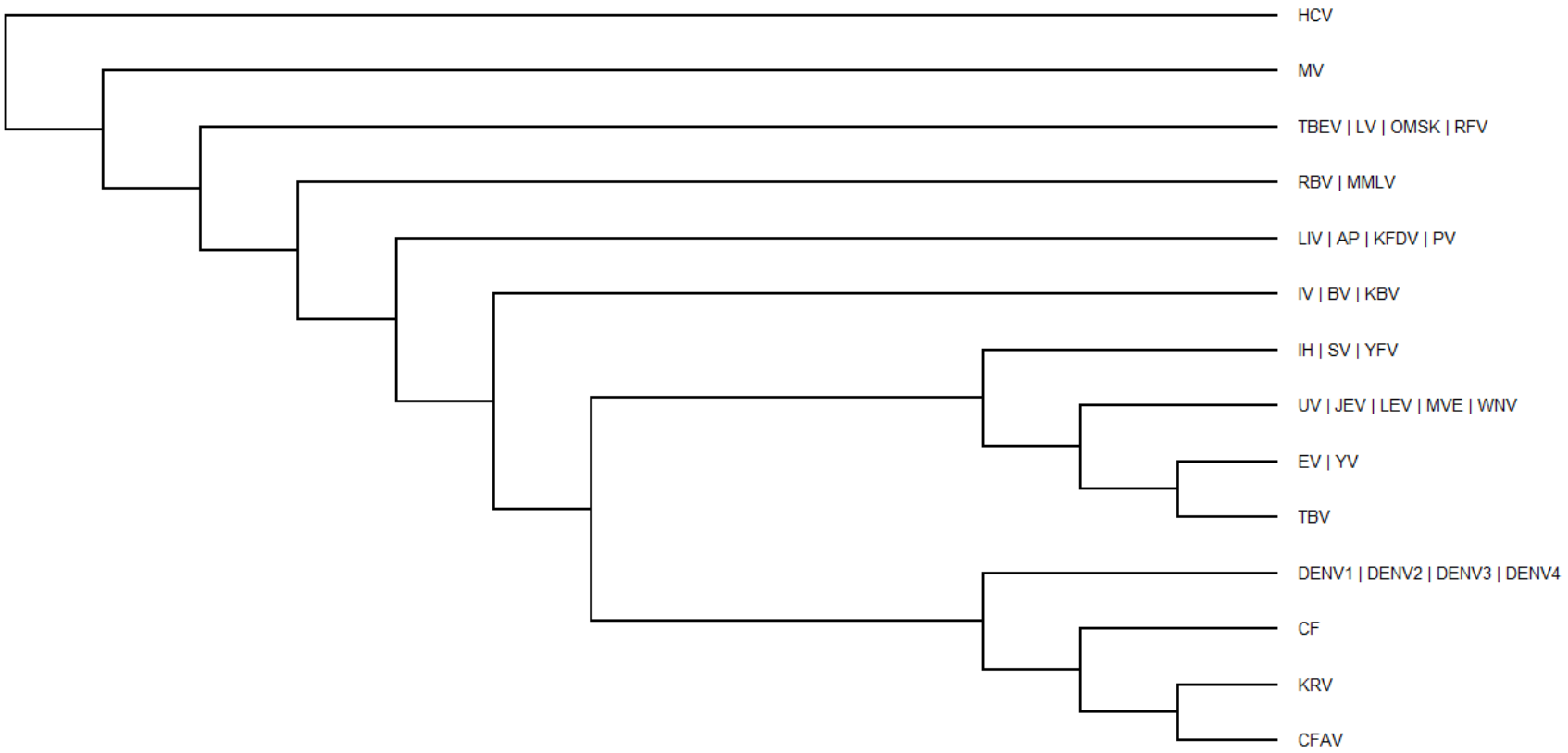
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{209}$PLSRNSTHEMYW$_{220}$



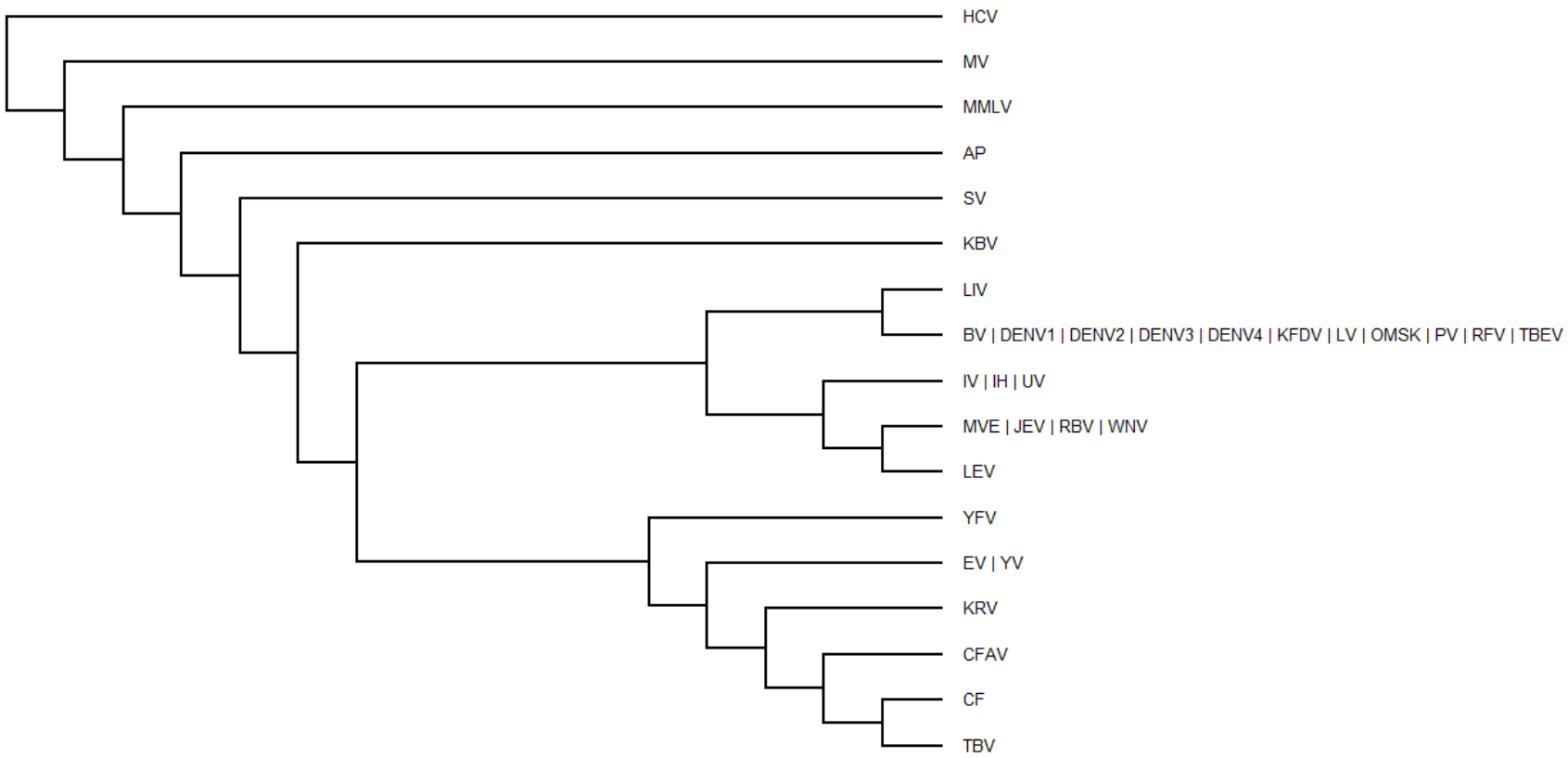
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{302}$TWAYHGSYE$_{310}$



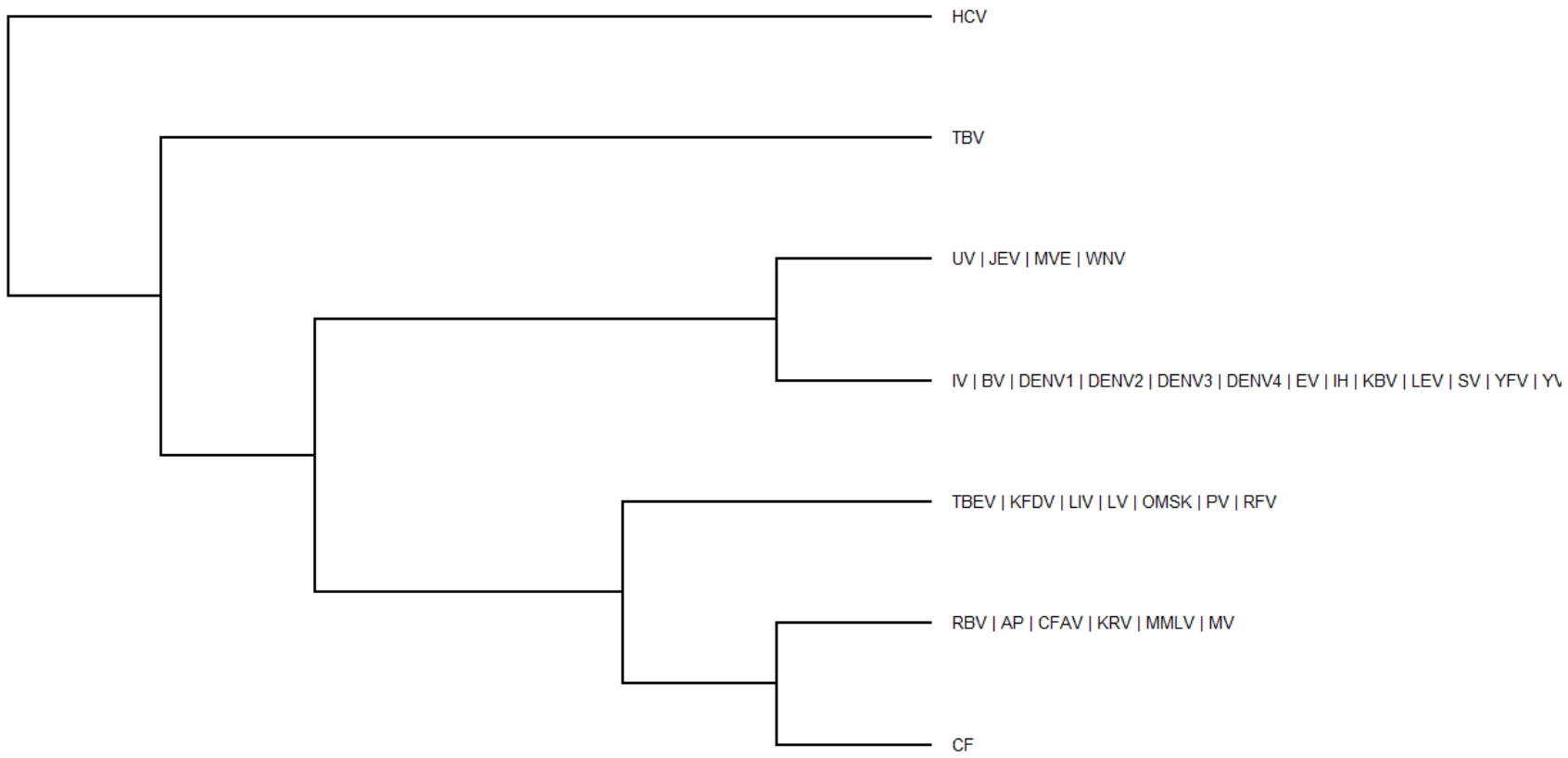
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{342}$AMTDTTPFGQQRVFKEKVDTRT$_{363}$



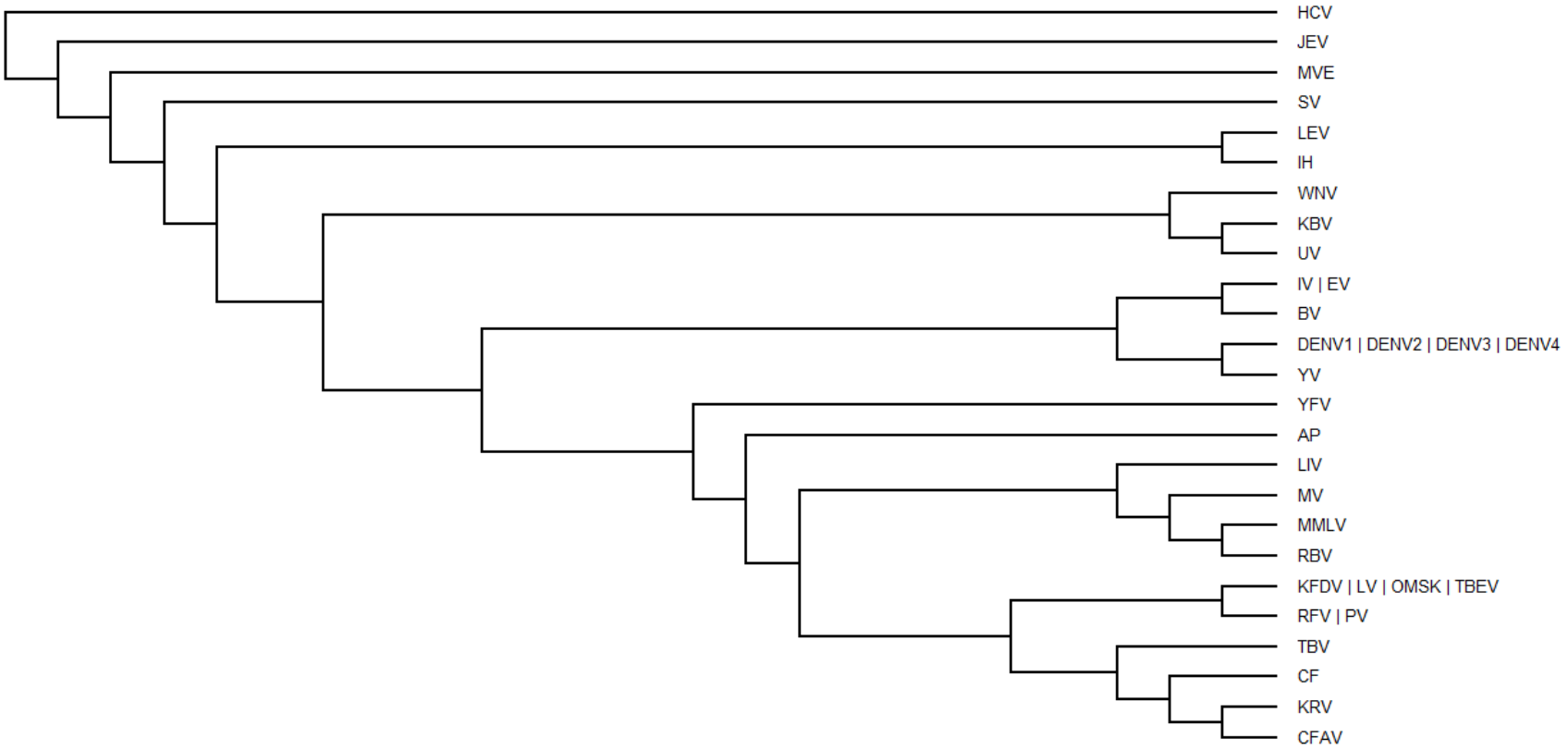
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence $NS5_{450}$CVYNMMGKREKKLGEFG$_{466}$



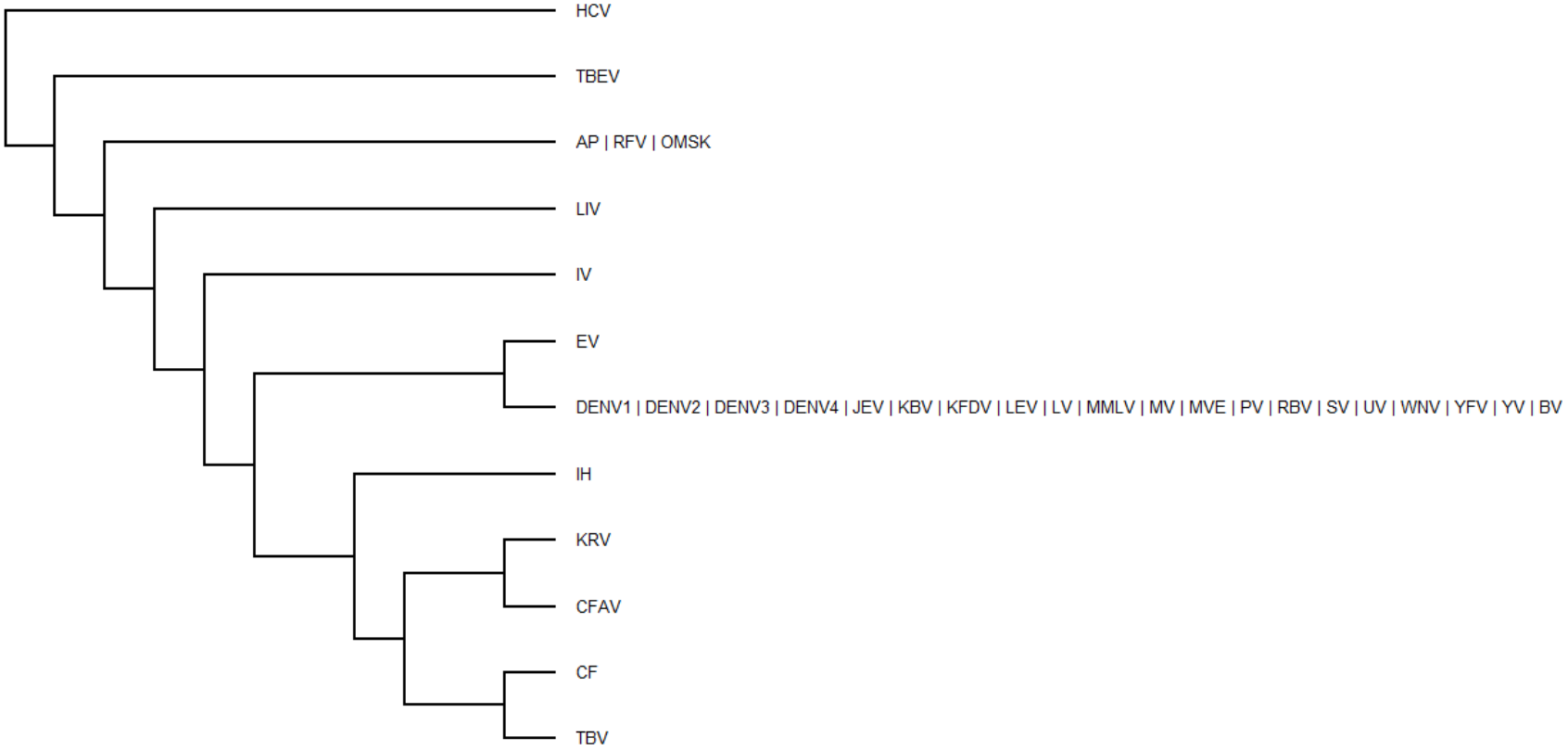
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{468}$AKGSRAIWYMWLGAR$_{482}$



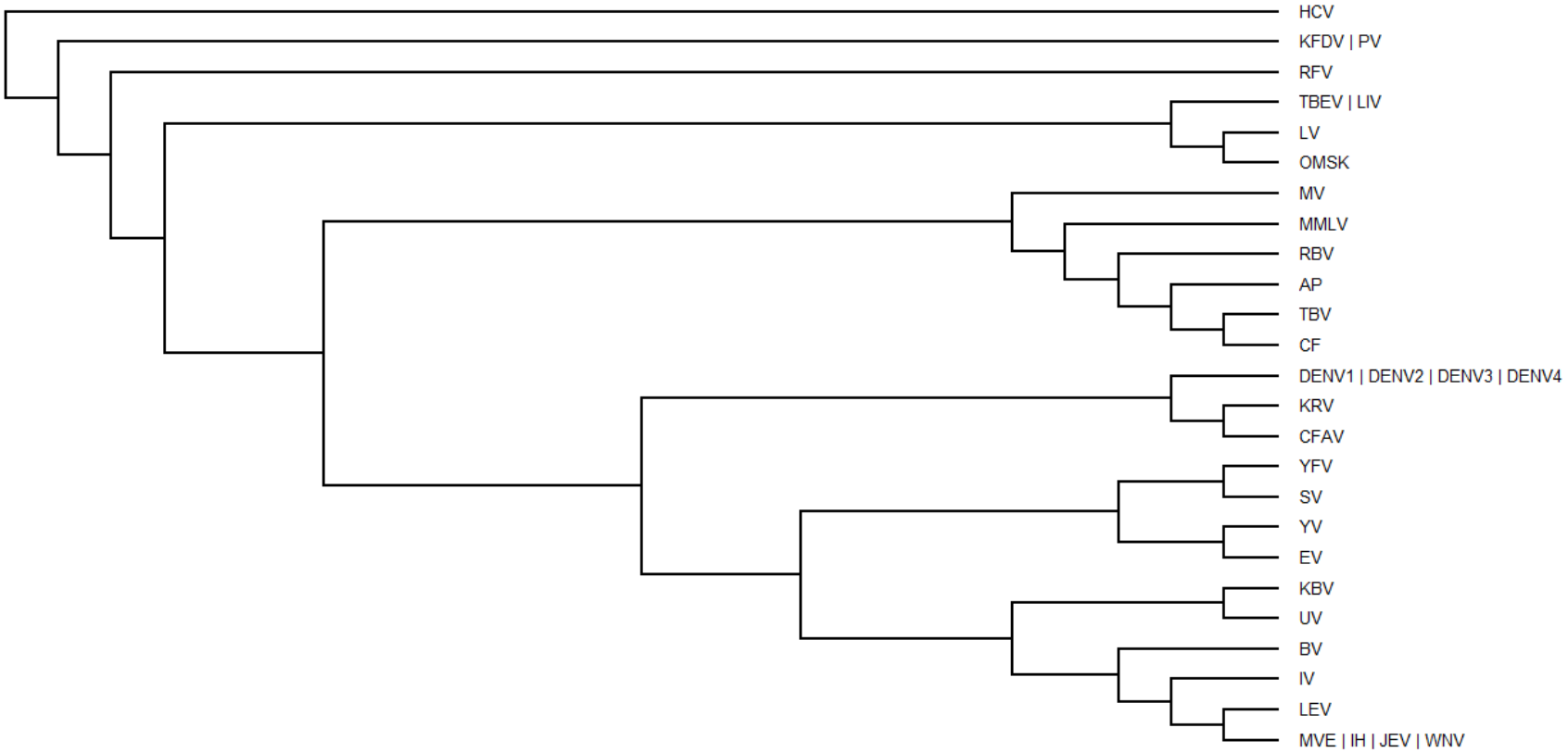
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{505}$SGVEGEGLH$_{513}$



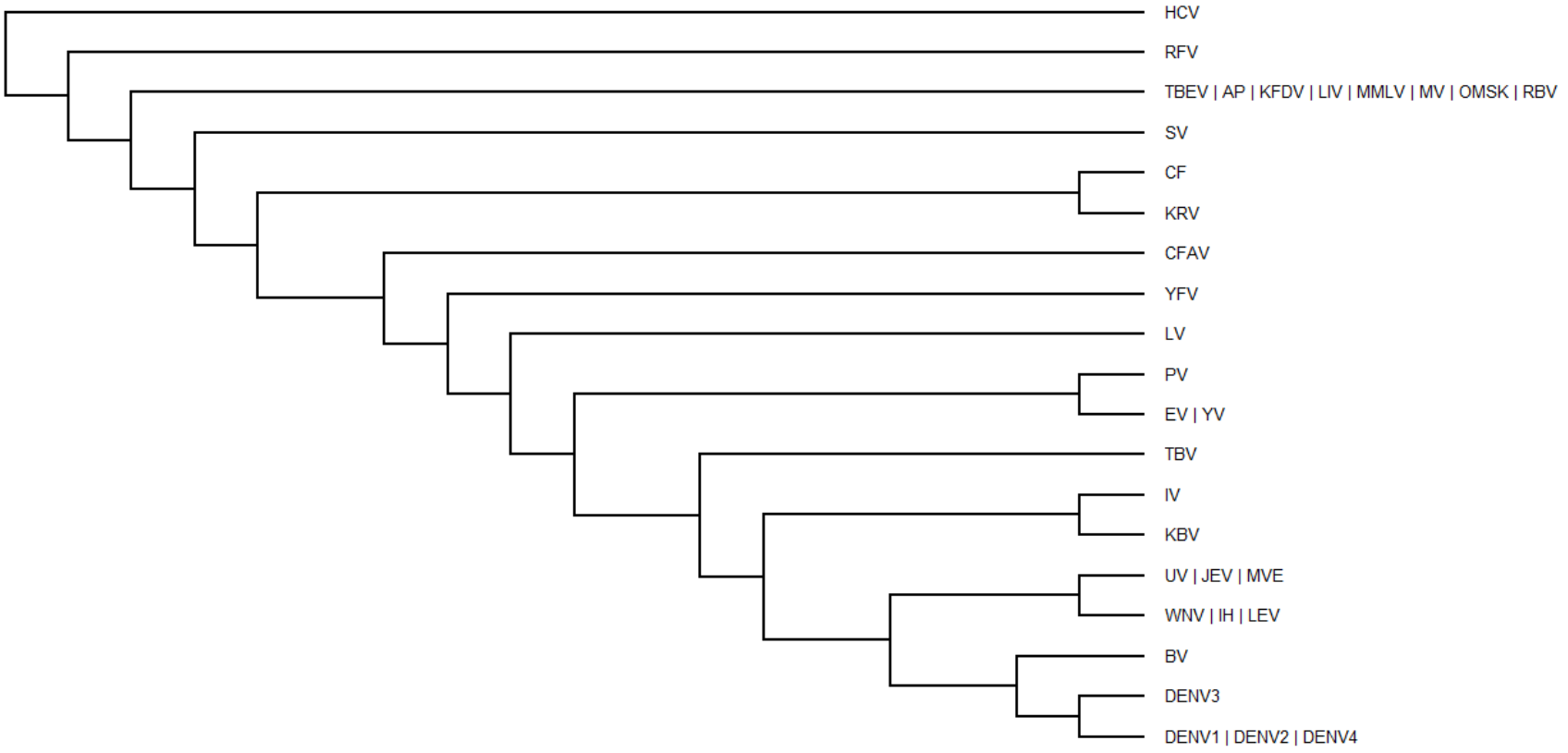
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{531}$YADDTAGWDTRIT$_{543}$



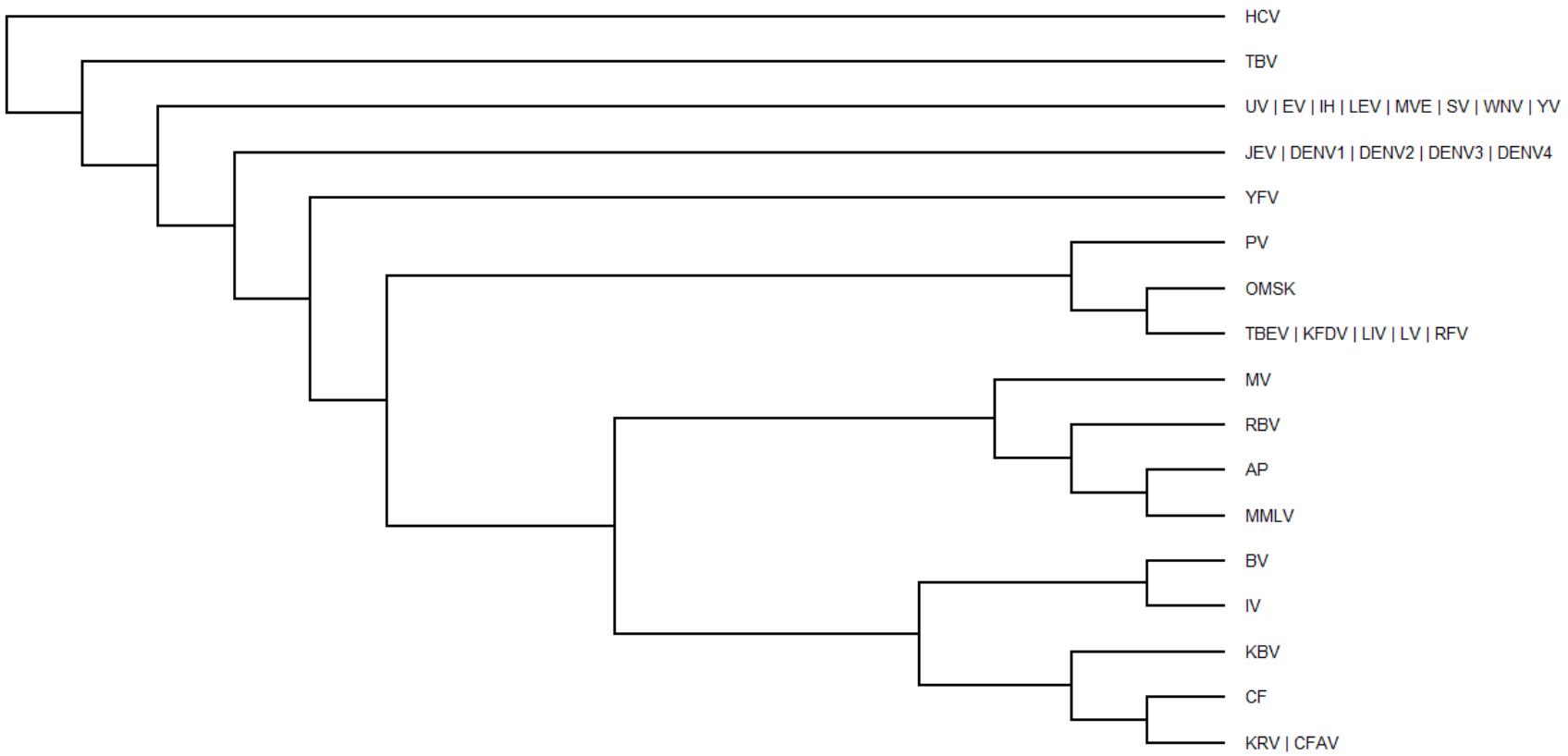
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{568}$IFKLTYQNKVV$_{578}$



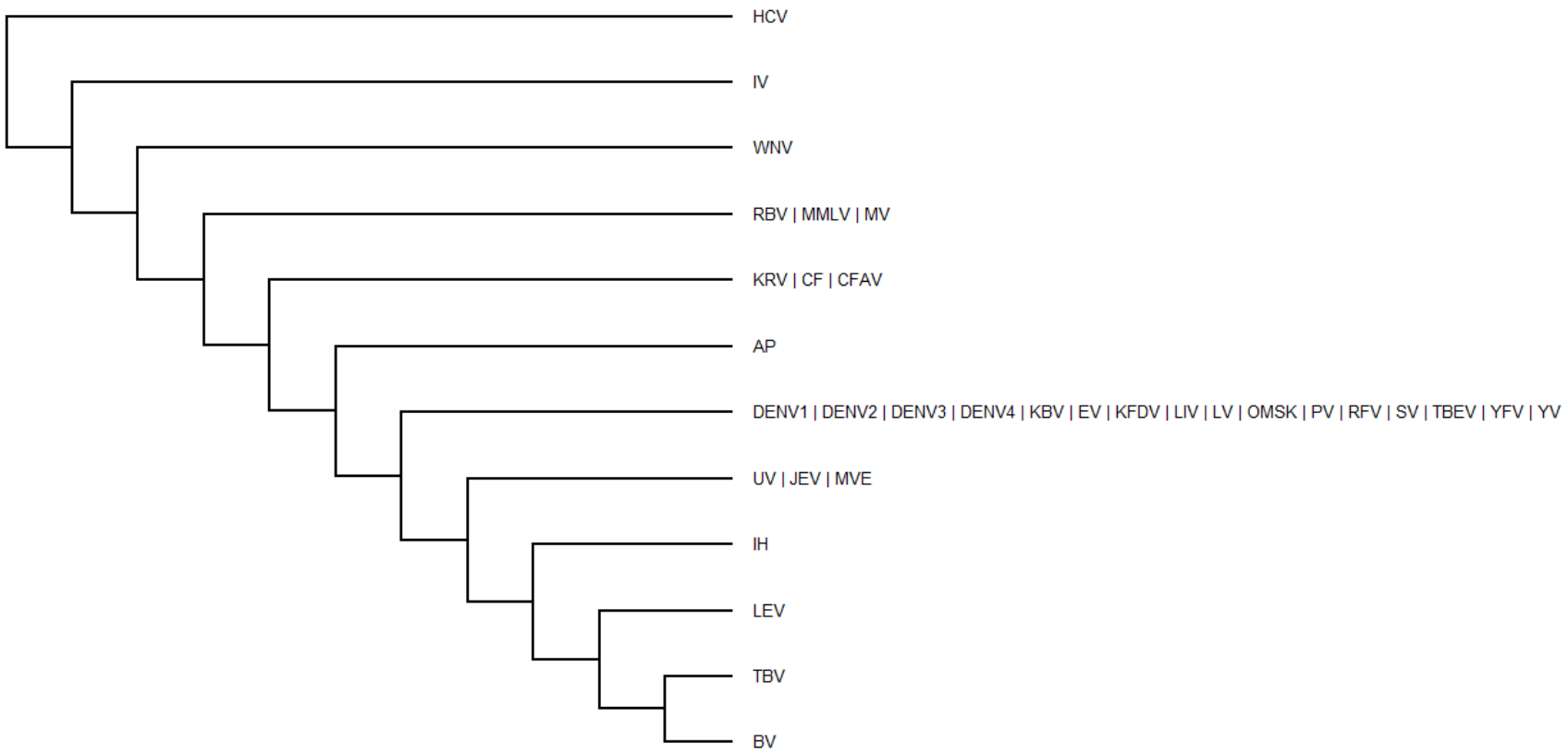
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{597}$DQRGSGQVGTYGLNTFTNME$_{616}$



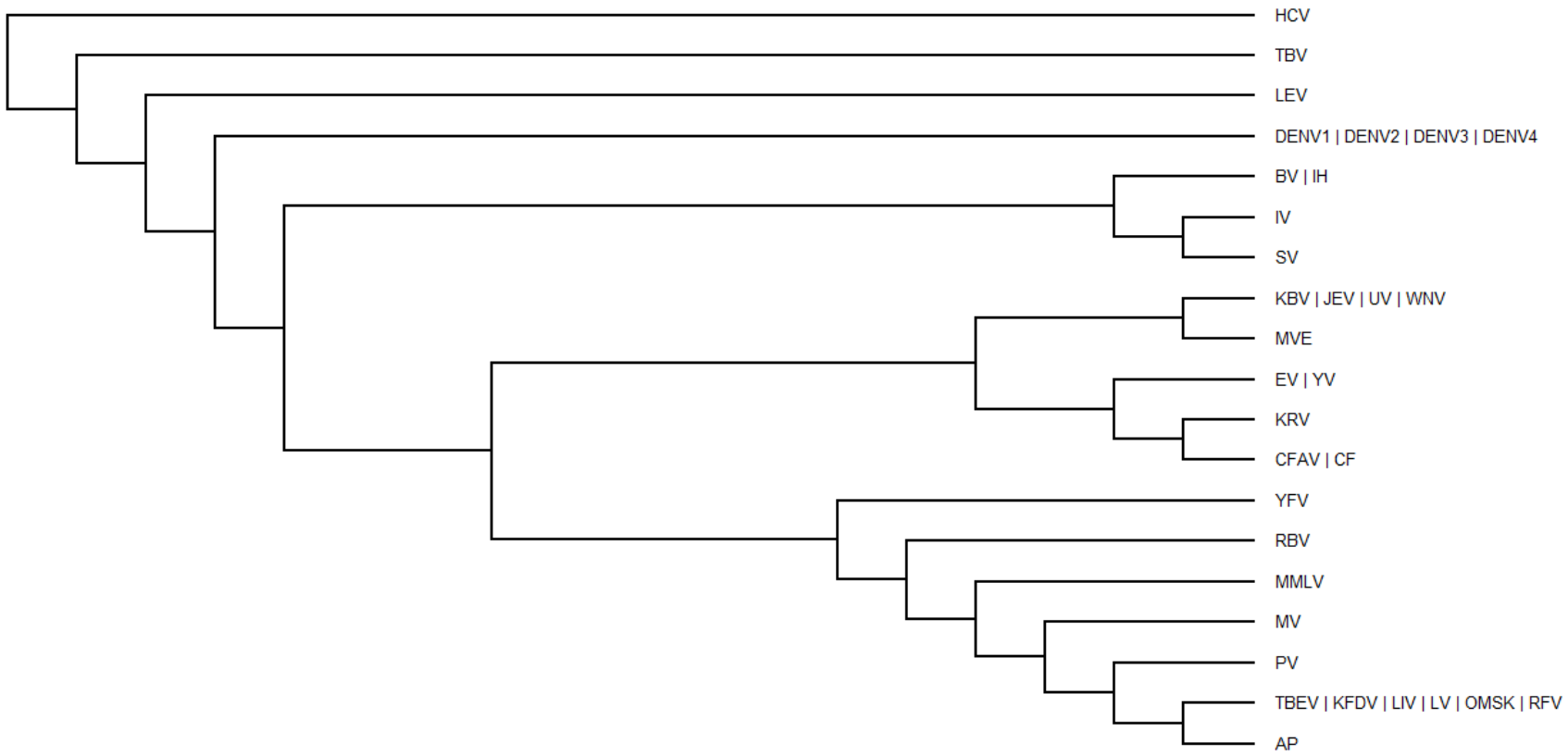
Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{658}$RMAISGDDCVVKP$_{670}$



Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{707}$VPFCSHHFH$_{715}$



Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{765}$LMYFHRRDLRLA$_{776}$



Evolution of sequences in selected *Flaviviruses* corresponding to the pan-DENV sequence NS5 $_{790}$PTSRTTWSIHA$_{800}$