

Conservation and Variability of West Nile Virus Proteins

Qi Ying Koo^{1,9}, Asif M. Khan^{1,9}, Keun-Ok Jung², Shweta Ramdas¹, Olivo Miotto^{1,3,4}, Tin Wee Tan¹, Vladimir Brusic⁵, Jerome Salmon², J. Thomas August^{2*}

1 Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **2** Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **3** MRC Centre for Genomics and Global Health, University of Oxford, Oxford, United Kingdom, **4** Mahidol-Oxford Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand, **5** Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America

Abstract

West Nile virus (WNV) has emerged globally as an increasingly important pathogen for humans and domestic animals. Studies of the evolutionary diversity of the virus over its known history will help to elucidate conserved sites, and characterize their correspondence to other pathogens and their relevance to the immune system. We describe a large-scale analysis of the entire WNV proteome, aimed at identifying and characterizing evolutionarily conserved amino acid sequences. This study, which used 2,746 WNV protein sequences collected from the NCBI GenPept database, focused on analysis of peptides of length 9 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 ($\leq 10\%$ of the WNV sequences analyzed). Eighty-eight fragments of length 9–29 amino acids, representing $\sim 34\%$ 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 contained experimentally confirmed human T-cell epitopes. This study identified a comprehensive catalogue of completely conserved WNV sequences, many of which are shared by other flaviviruses, and majority are potential epitopes. 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.

Citation: Koo QY, Khan AM, Jung K.-O., Ramdas S, Miotto O, et al. (2009) Conservation and Variability of West Nile Virus Proteins. PLoS ONE 4(4): e5352. doi:10.1371/journal.pone.0005352

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received: February 2, 2009; **Accepted:** March 10, 2009; **Published:** April 29, 2009

Copyright: © 2009 Koo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This project has been funded in part with the Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, USA, under Contract No. HHSN266200400085C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: taugust@jhmi.edu

9 These authors contributed equally to this work.

Introduction

West Nile virus (WNV) is a mosquito-borne pathogen of the family *Flaviviridae*, genus *Flavivirus*, closely related to other important human pathogens, such as yellow fever (YFV), Japanese encephalitis (JEV), and dengue (DENV) viruses, among others. The genome is a single-stranded positive-sense RNA encoding a polyprotein precursor of approximately 3,430 amino acids, which is cleaved into three structural (capsid, C; precursor membrane and membrane, prM/M; envelope, E) and seven nonstructural proteins (NS1, 2a, 2b, 3, 4a, 4b and 5) [1,2]. WNV is present predominantly amongst avian hosts, and can infect humans through incidental zoonotic transmission via mosquitoes [3]. The virus is endemic in many parts of Africa, Asia, Europe, and most recently North America [4]. At present, there is no registered human vaccine or specific therapy to prevent or treat WNV infection. Although the majority of infected humans remain asymptomatic, about 20% experience influenza-like symptoms, and approximately 1 in 150 develop severe illness, including meningoencephalitis [5,6].

Like other RNA viruses, WNV exhibits significant genetic diversity, a consequence of mainly high mutation rates in RNA replication, and subsequent selection of mutants adapted to changing environment [7,8]. Five distinct genotypes have been identified by phylogenetic analyses of the C-prM-E region, differing from each other by 20–25% across the complete genome [9]. However, variability is uneven across the viral genome, since mutations detrimental to viral fitness are restricted. Thus, while certain protein sites permit multiple mutations, sites critical to viral structure-function are evolutionarily robust and highly conserved. The analysis of the evolutionary dynamics and immunogenic properties of these sites has relevance to multiple applications, including the design of diagnosis, drugs and vaccines.

The focus of this study is to identify and characterize WNV protein regions that have exhibited strong conservation throughout the recorded history of the virus, and that are potential targets of T-cell immune responses. T-cell responses have been implicated in the control and clearance of WNV infection [10–15]. Their specificities are governed by human leukocyte antigen (HLA)

binding to peptides, derived from proteolysis of antigen (Ag) proteins, for presentation to T cells. HLA class I and class II molecules present peptides to CD8⁺ and CD4⁺ T-cells respectively, which play a critical role in cytotoxic responses and in the induction and maintenance of Ag-specific memory responses.

A bioinformatics approach is applied herein to (a) examine the large number of WNV sequences available in public databases, (b) analyze the conservation and variability of these sequences, (c) identify sequence fragments of WNV proteins that are completely conserved in all known WNVs (henceforth referred to as pan-WNV sequences), (d) examine the structure-function relationship and distribution in nature of pan-WNV sequences, and (e) assess the immune relevance of pan-WNV sequences as potential T-cell epitopes, correlating immunoinformatic predictions to previously reported human WNV T-cell epitopes and to our current studies in the identification of human WNV T-cell epitopes by use of HLA transgenic mice.

Methods

Methodology overview

The overall bioinformatics approach to this study is summarized in **Figure 1**. The rationale for this approach and methodology is consistent with that of previous studies [16–18], and is briefly described below.

Data preparation, selection and alignment

WNV sequence records were retrieved from the NCBI Entrez protein database [19] in June 2007 by searching the NCBI taxonomy browser for WNV (taxonomy ID 11082).

The sequences of the 10 WNV proteins (C, prM, E, NS1, NS2a, NS2b, NS3, NS4a, NS4b and NS5) were extracted from the downloaded records by performing BLAST [20] against the record sequences using the individual protein sequences as queries, obtained from the annotated WNV reference record P06935 (Swiss-Prot/TrEMBL [21]). Multiple sequence alignment of extracted sequences for each viral protein was performed by use of MUSCLE v3.6 [22]. All multiple sequence alignments were manually inspected and corrected wherever necessary. All extracted WNV protein sequences,

whether partial or full-length, were included in all analyses, 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. We retained the duplicate sequences (2,206) for our 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.

Amino acid difference between WNV protein sequences

Pair-wise percentage amino acid difference for the full-length unique sequences of each WNV proteins was computed by use of ClustalW 1.83 [23] with default parameters. This was done to survey the extent of amino acid variation in the WNV data of 2007.

Nonamer entropy analysis of WNV sequences

Entropy analysis [24,25] was carried out as described in [17], by use of the Antigenic Variability Analyser tool (AVANA), to study the diversity of WNV protein sequences over the period which the sequences were collected. Entropy measurements were based on 9 amino acid peptides (nonamers), since it is the typical length of epitopes that bind to HLA class I molecules, and of the binding cores of HLA class II epitopes [26]. At any given position x in the alignment, nonamer peptide entropy $H(x)$ was calculated by

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

where $p_{i,x}$ represents the measured probability of a particular nonamer peptide i occurring with its center position at x ; $n(x)$ represents the total number of peptides observed at position x – larger number of peptides generally result in increased entropy values. Positions with highly dominant (conserved) peptides are characterized

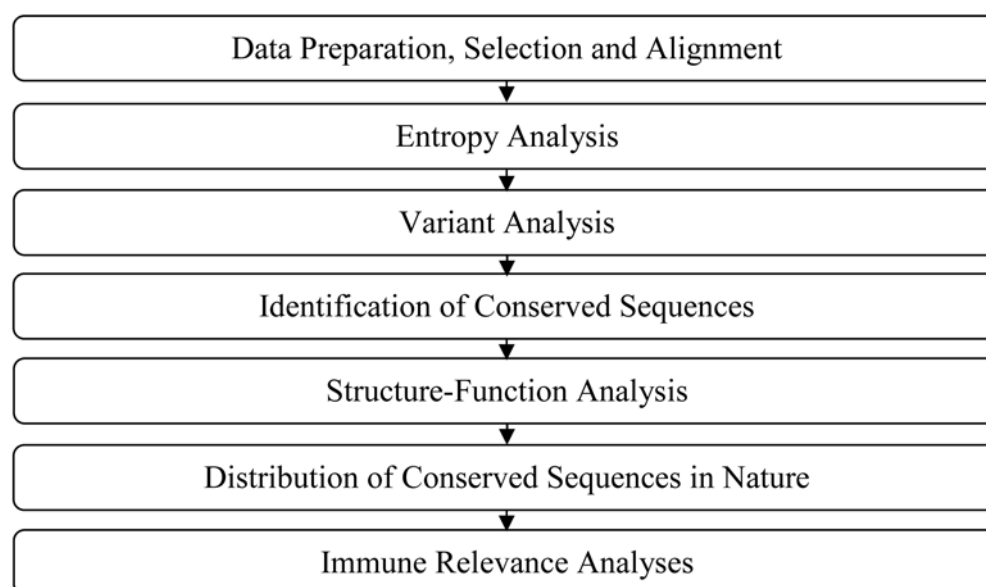


Figure 1. General overview of the bioinformatics approach employed in this study.

doi:10.1371/journal.pone.0005352.g001

by low entropy values, approaching zero. In the case of incomplete sequences, only sequences with a valid nonamer centered at position x were included in the computation, and positions where more than 50% of sequences contained gaps were discarded for their low statistical significance. 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. For sequence sets of finite size, entropy calculations are affected by size bias, which were corrected by a statistical sub-sampling method, as described in [17].

Nonamer variant analysis of WNV sequences

Variable amino acid sites were further analyzed by computing the representation of nonamer variants. At any given position x , variant nonamers were defined as nonamers that differed by at least one amino acid from the predominant nonamer (the most common peptide) at that position. Further details of this analysis are given in [17].

Identification of completely conserved WNV sequences

Completely conserved sequences (pan-WNV sequences) of at least 9 amino acids and fully identical in all the sequences analyzed (100% representation) were identified from the multiple sequence alignment of each protein. Peptides containing unknown residues (X) were ignored.

Structure-function analysis of pan-WNV sequences

A literature search was conducted to identify reported and putative functional properties of the pan-WNV sequences, including a search of the Prosite database [27] using ScanProsite [28], and a search of the Pfam database [29]. The conserved sequences were also mapped onto the three-dimensional (3-D) structures of WNV proteins whenever these were available in the protein data bank (PDB) [30]. Only 3-D structures obtained via X-ray diffraction were utilized for mapping, and were visualized by use of the CPK representation in the ICM Browser v3.5 (www.molsoft.com).

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

Pan-WNV sequences with at least 9 consecutive amino acids in common with other viruses were identified by performing BLAST search against protein sequences of all viruses (txid10239) reported at NCBI (as of August 2007), except WNV (txid11082): parameters included search by Entrez query limited to "txid10239[Organism:exp] NOT txid11082[Organism:exp]", "automatically adjust parameters for short sequences" option disabled, "low-complexity" filter disabled, maximum number of aligned sequences to be displayed set to "20,000", expect threshold set to "200,000", or "20,000", or "2,000" until a valid result was obtained, word size set to "2", matrix set to "PAM30", gap costs set to "Existence: 9, Extension: 1", compositional adjustments set to "no adjustment". Similar BLAST searches were carried out against protein sequences of all organisms, except viruses: parameters were the same as the previous search against all viruses, excluding WNV, except that the search by Entrez query was limited to "Root[ORGN] NOT Viruses[ORGN] NOT txid81077[ORGN]". Artificial sequence hits were removed by the "NOT txid81077[ORGN]" keyword.

Identification of known and predicted WNV HLA-supertype binding epitopes

A literature search and a search of the Immune Epitope Database [31] (www.immuneepitope.org) identified previously

reported HLA class I and II human T-cell epitopes of WNV that overlapped at least 9 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. Putative HLA class I supertype-restricted peptides were predicted by use of NetCTL [32], Multipred [33], ARB [34], and class II supertype-restricted peptides by Multipred and TEPITOPE [35], following the specifications as described in [17].

Results

WNV protein sequence datasets

A total of 2,746 complete and partial WNV protein sequences were extracted from the NCBI Entrez protein database records as of June 2007 (**Table 1; Data Set S1**). The large number of sequences analyzed and their wide spatial and temporal (1955–2005; based on information available in annotated NCBI records) distribution enabled a broad survey of WNV protein 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 variation 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 1**).

Evolutionary stability of WNV

The evolutionary diversity of WNV was studied by computing entropy as described in the Methods. The entropy plot revealed the evolutionary variability of nonamer sequences across the WNV proteome (**Figure 2**). 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,

Table 1. Number of WNV protein sequences retrieved from NCBI and their maximum percentage amino-acid difference.

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	-

^aRetrieved from NCBI Entrez Protein Database on 28th June 2007.

^bApproximate size indicated in number of amino acids.

^cMaximum percentage amino-acid difference for each WNV protein, computed using ClustalW [23].

doi:10.1371/journal.pone.0005352.t001

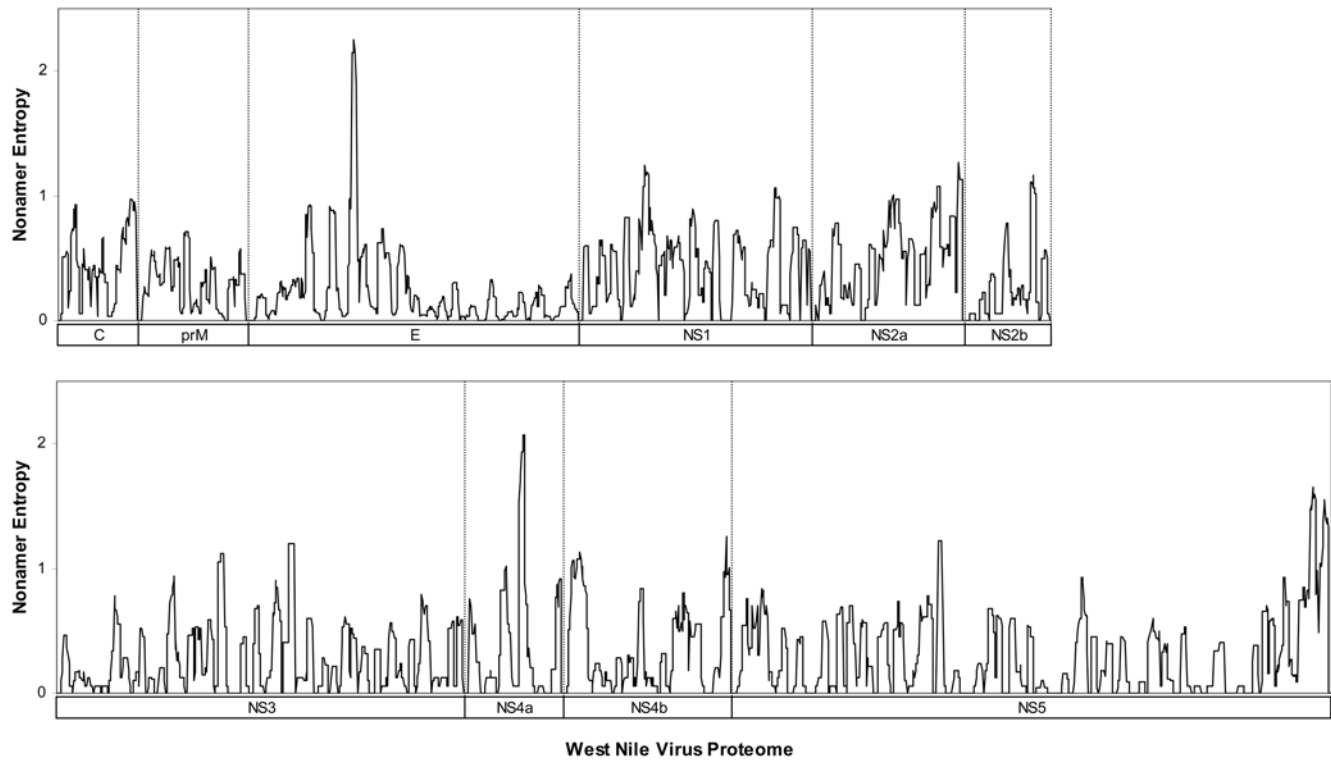


Figure 2. Peptide entropy plots for WNV protein alignments.
doi:10.1371/journal.pone.0005352.g002

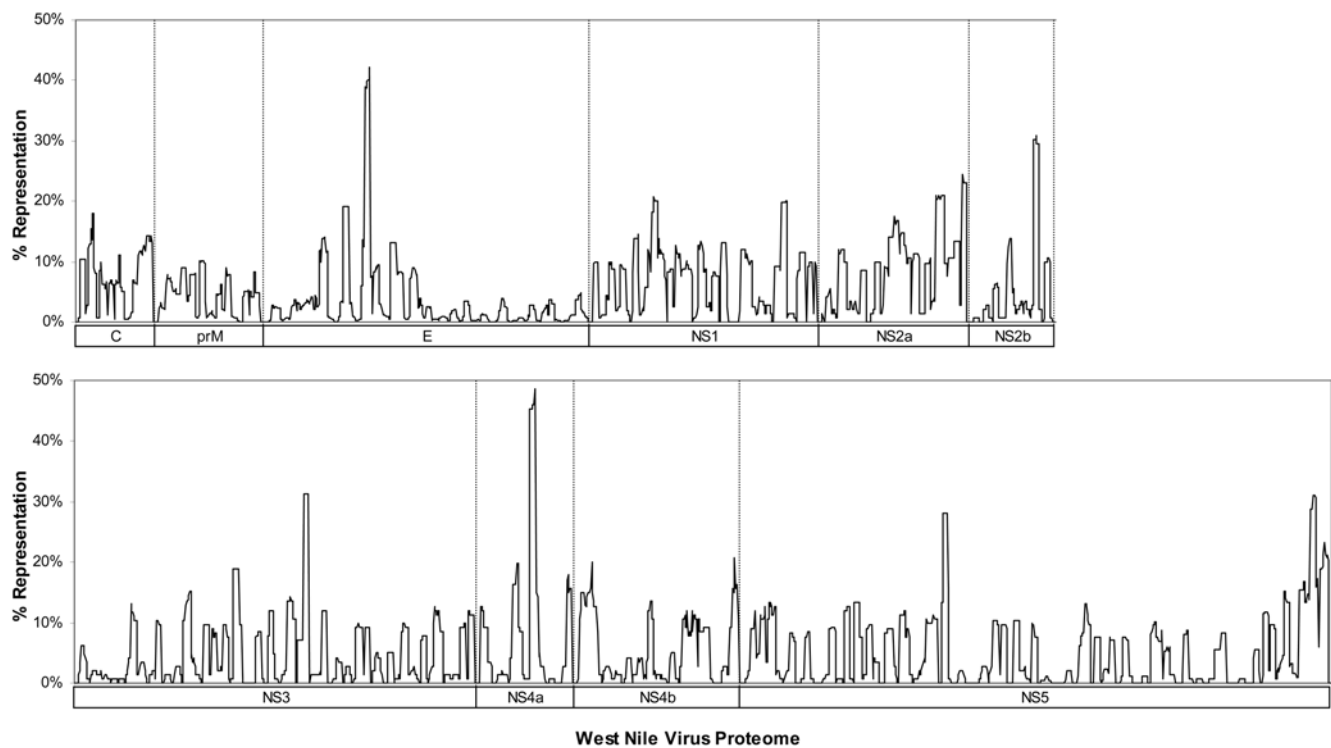


Figure 3. Percentage representation of nonamer variants in relation to the predominant nonamer peptide for all nonamer positions in WNV protein alignments.
doi:10.1371/journal.pone.0005352.g003

Table 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-LLLLVPAYS-167
E	11	1-FNCLGMSNRDF-11
	14	104-GCGLFGKGSIDTCA-117
	9	293-LKGTTYGVC-301
	19	338-SVASLNDLTPVGRLVTNP-356
	12	370-ELEPPFGDSYIV-381
	10	417-LGDTAWDFGS-426
	12	449-LFGGMSWITQGL-460
	11	58-RSVSRLEHQMW-68
NS1	9	114-GWKAWGKSI-122
	9	154-EVEDFGFGL-162
	10	195-HSDLSYWIES-204
	25	209-TWKLERAVLGEVKSCTWPETHLWG-233
	11	276-DFDYCPGTTVT-286
	10	313-CRSCTLPPLR-322
	10	328-GCWYGMEIRP-337
	NS2a	10
15		69-NSGGDVVHLALMATF-83
NS2b	10	1-GWPATEVMTA-10
	14	12-GLMFAIVGGLAELD-25
	9	32-PMTIAGLMF-40
NS3	11	108-SAYTPWAILPS-118
	10	1-GGVLWDTSP-10
	10	20-TGVYRIMTRG-29
	10	52-TTKGAALMSG-61
	10	63-GRLDPYWGSV-72
	10	74-EDRLCYGGPW-83
	12	108-NVQTKPGVFKTP-119
	12	131-PTGTSGSPIVDK-142
	13	145-DVIGLYNGVIMP-157
	11	161-YISAIQGERM-171
	12	191-VLDLHPGAGKTR-202
	9	235-ALRGLPIRY-243
	27	256-EIVDVMCHATHRLMSPHRVPNPNLF-282
	14	288-HFTDPASIAARGYI-301
	12	310-AAAFMTATPPG-321
	10	337-QTEIPDRAWN-346
	10	357-GKTWVFPVPSV-366
16	385-QLNRKSYETEPKCKN-400	
11	408-TTIDISEMGANF-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	

Table 2. cont.

WNV protein	Length (aa)	Pan-WNV sequence ^a
	10	563-WCFDGPRTNT-572
NS4a	15	19-KTWEALDTMYVVATA-33
	11	43-ALEELPDALQT-53
	13	101-GTKIAGMLLLSLL-113
	20	115-MIVLIPEPEKQRSQTDNQLA-134
NS4b	9	39-PATAWSLYA-47
	13	68-TSLTSINQASAL-80
	9	85-RGFPFVDVG-93
	12	138-AQRRTAAGIMKN-149
	10	156-VATDVPELER-165
	22	208-VTLWENGASSVWNATTAIGLCH-229
NS5	10	1-GGAKGRTLGE-10
	9	60-AKLRWLVER-68
	17	79-DLGCGRGGWCYMATQK-95
	22	107-GPGHEEPQLVQSYGWNIVTMKS-128
	10	141-DTLCDIGES-150
	10	152-SSAEVEEHRT-161
	9	168-VEDWLHRGP-176
	16	208-RNPLSRNSTHEMYWVS-223
	12	235-MTSQVLLGRMEK-246
	10	259-NLGSSTRAVG-268
	13	299-NHPYRTWNYHGSY-311
	18	318-SASSLVNGVRLLSKPWD-335
	29	340-VTTMAMTDTPFGQQRVFKEKVDKAPEP-368
	10	375-VLNETTNWLW-384
	18	404-KVNSNAALGAMFEEQNQW-421
	10	440-EREHLRGEC-449
	12	451-TCIYNNMMGKREK-462
29	472-GSRAIWFMWLGARFLEFEALGFLNEDHWL-500	
16	504-NSGGGVEGLGLQKLG-519	
13	533-YADDTAGWDTRIT-545	
10	548-DLENAKVL-557	
15	571-IELTYRHKVVKMMP-585	
23	596-ISREDQRGSGQVVTYALNTFTNL-618	
12	620-VQLVRMMEGEGV-631	
19	662-RMAVSGDDCVKPLDDRFA-680	
14	689-MSKVRKDIQEWKPS-702	
18	704-GWYDWWQVPCSNHFTL-721	
27	741-GRARISPGAGWNRDTACLAKSYAQMW-767	
21	769-LLYFHRRDLRLMANAICSAVP-789	
12	792-WVPTGRTTWSIH-803	

^aNumbers prefixing and affixing sequences represent start and end positions in the protein alignment.

doi:10.1371/journal.pone.0005352.t002

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.

Representation of variant WNV sequences

Completely conserved nonamer sites with zero variant 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 3**). The position with the highest representation of variant nonamer sequences (49%) was found in the nonstructural protein NS4a. Overall, our data suggest a low probability of immune challenges from variant WNV T-cell epitopes, due to a high representation of historically conserved sequences of the WNV proteome in the known virus data.

Completely conserved pan-WNV sequences

A total of 88 completely conserved sequence fragments (pan-WNV sequences) were identified across the whole proteome (**Table 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 (3,430 aa) (**Table 3**). The C protein had no completely conserved nonamer fragment, which is consistent with the large number of amino acid difference (23%) observed for sequences of this protein, compared to other WNV proteins (**Table 1**). The NS3 and NS5 proteins contained the greatest number of completely conserved fragments, 25 in NS3 (spanning 48% of the protein length) and 30 in NS5 (spanning 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 completely conserved sequences (14% of the protein length), while E had 7 (18% of the protein length).

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 [36]. A search in the Prosite and Pfam

databases [27,29] revealed that 50 of the 88 pan-WNV sequences are known to be associated with putative or known biological functions and/or structure (**Table 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 [37], while two correspond to immunoglobulin-like domain, attributed to putative receptor binding sites [38]. One NS1 sequence correspond to the putative ATP/GTP binding site p-loop motif, likely to be involved in helicase activity [39]. NS3 contained 4 pan-WNV sequences that correspond to the peptidase family S7 (*Flavivirus* serine protease) domain [40], and 4 that correspond to known/putative *Flavivirus* Asp-Glu-Ala-Asp/His (DEAD/H) domain associated with ATP-dependent helicase activity [41]. NS5 contained 17 sequences that correspond to the RNA-dependent RNA polymerase (RdRp)/catalytic domain [42,43]. 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) (**Figure S1**). 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.

Distribution of pan-WNV sequences in nature

Sixty-seven (67) of the 88 pan-WNV sequences (~76%) overlapped at least 9 amino acid sequences of as many as 68 other viruses of the family *Flaviviridae*, genus *Flavivirus* (**Figure 4**). Each of these 67 sequences matched at least one, and at most 61 *Flavivirus* species (**Figure 5 and Table S1**). Murray valley encephalitis (MVE) virus shared 49 of the 67 pan-WNV

Table 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

^aApproximate length indicated in number of amino acids, according to the reference protein sequence described in the *Methods*.

^bApproximate percentage rounded off to nearest whole number.

doi:10.1371/journal.pone.0005352.t003

Table 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-SVASLNDLTPVGRVTVNP-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-TWKLERAVLGEVKSCTWPETHLWG-233	-	PKC, CKII	
	328-GCWYGMIEIRP-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-PTGTSGSPVDK-142	Peptidase S7	-	
	145-DVIGLYNGVIMP-157	Peptidase S7	N-myristoylation	
	191-VLDLHPGAGKTR-202	DEAD/H domain	N-myristoylation	
	256-EIVDVMCHATLTHRLMSPHRVNPYNLF-282	DEAD/H domain	PKC	
	288-HFTDPASIAARGYI-301	DEAD/H domain	-	
	310-AAAFMTATPPG-321	DEAD/H domain	-	
	385-QLNRKSYETEPKCKN-400	-	TK	
	422-RVIDSRKSVKP-432	-	PKC	
	470-GDEYCYGGHTNEDDSN-485	-	CKII, N-myristoylation	
	NS4a	101-GTKIAGMLLSLL-113	-	N-myristoylation
		115-MIVLIPEKQRSQTDNQLA-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-SASSLVNGVRLSKPWD-335	RdRp	-	
	340-VTTMAMTDTPFGQQRVFEKVDKAPPEP-368	RdRp	-	
	375-VLNETTNWLW-384	-	N-glycosylation	
	404-KVNSNAALGAMFEEQNQW-421	RdRp	-	
	451-TCIYNMMGKREK-462	RdRp	Amidation	
	472-GSRAIWMWLGARFLEFEALGFLNEDHWL-500	RdRp	-	
	504-NSGGGVEGLGLQKLG-519	RdRp	N-myristoylation	
	533-YADDTAGWDTRIT-545	RdRp	-	
	571-IELTYRHKVVKVMP-585	RdRp	PKC	
	596-ISREDQRGSGQVVTYALNTFTNL-618	RdRp/ RdRp catalytic domain	CKII, N-myristoylation	
	620-VQLVRMEGEGV-631	RdRp	-	
662-RMAVSGDDCVVKPLDDRFA-680	RdRp/ RdRp catalytic domain	CKII		
689-MSKVRKDIQEWKPS-702	RdRp	-		
704-GWYDQQVFPFCSNHFTL-721	RdRp	-		

Table 4. Reported biological properties of pan-WNV sequences.

WNV protein	Pan-WNV sequence	Functional domains and motifs ^a	Putative post-transcriptional modifications ^a
	741-GRARISPGAGWNVDRDTACLAKSYAQMW-767	RdRp	N-myristoylation
	769-LLYFHRRDLRLMANAICSAVP-789	RdRp	-
	792-WVPTGRTTWSIH-803	RdRp	PKC

^aProsite (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.
doi:10.1371/journal.pone.0005352.t004

sequences; JEV and Usutu viruses shared 47 and 41, respectively; and representatives of some of the important human pathogens, St. Louis encephalitis (LEV), DENV, YFV, and tick-borne encephalitis (TBEV) viruses shared from 36 to 11 of the 67 pan-WNV sequences. The representation (frequency) of these pan-WNV sequences ranged from low to high across reported sequences of the several well studied flaviviruses, including DENV, JEV, Louping ill (LIV), Omsk hemorrhagic fever (OMSK), Powassan (PV), LEV, TBEV, and YFV (**Table S1**). For example, the pan-WNV sequence E₁₀₄₋₁₁₇ was present in 99% of the 245 E protein JEV sequences, while E₂₉₃₋₃₀₁ 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 sequences 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, 5 of the 88 pan-WNV sequences (prM₁₅₈₋₁₆₇, NS3₄₀₈₋₄₁₈, NS4b₂₀₈₋₂₂₉, NS5₁₋₁₀, and NS5₅₀₄₋₅₁₉) shared 9 consecutive amino acids with 7 non-viral species. The nonamer sequence from prM₁₅₈₋₁₆₆ is found in the bacterium *Acidiphilium cryptum* JF-5; NS3₄₀₉₋₄₁₇ in the mosquito *Aedes albopictus*; NS4b₂₁₈₋₂₂₆ in the Japanese rice *Oryza sativa* (japonica cultivar-group); NS5₂₋₁₀ in the bacterium *Actinomyces odontolyticus*, NS5₅₀₄₋₅₁₂ in the bacteria *Burkholderia ambifaria* MC40-6 and *Burkholderia cepacia* AMMD; and NS5₅₀₆₋₅₁₄ in the bacterium *Methylobacterium extorquens* PA1.

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

A literature survey and IEDB database search revealed that 3 of the pan-WNV sequences (2 in NS3, and one in NS5) matched 3

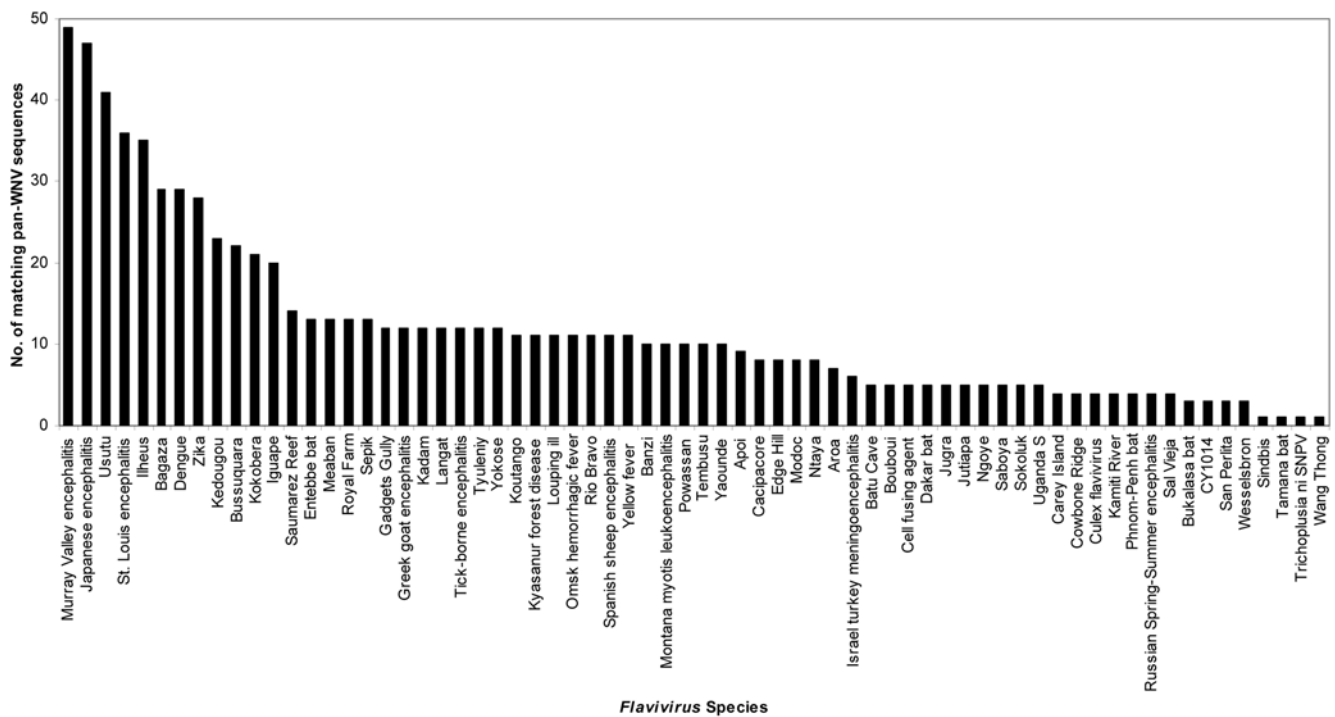


Figure 4. Number of pan-WNV sequences conserved in other flaviviruses.
doi:10.1371/journal.pone.0005352.g004

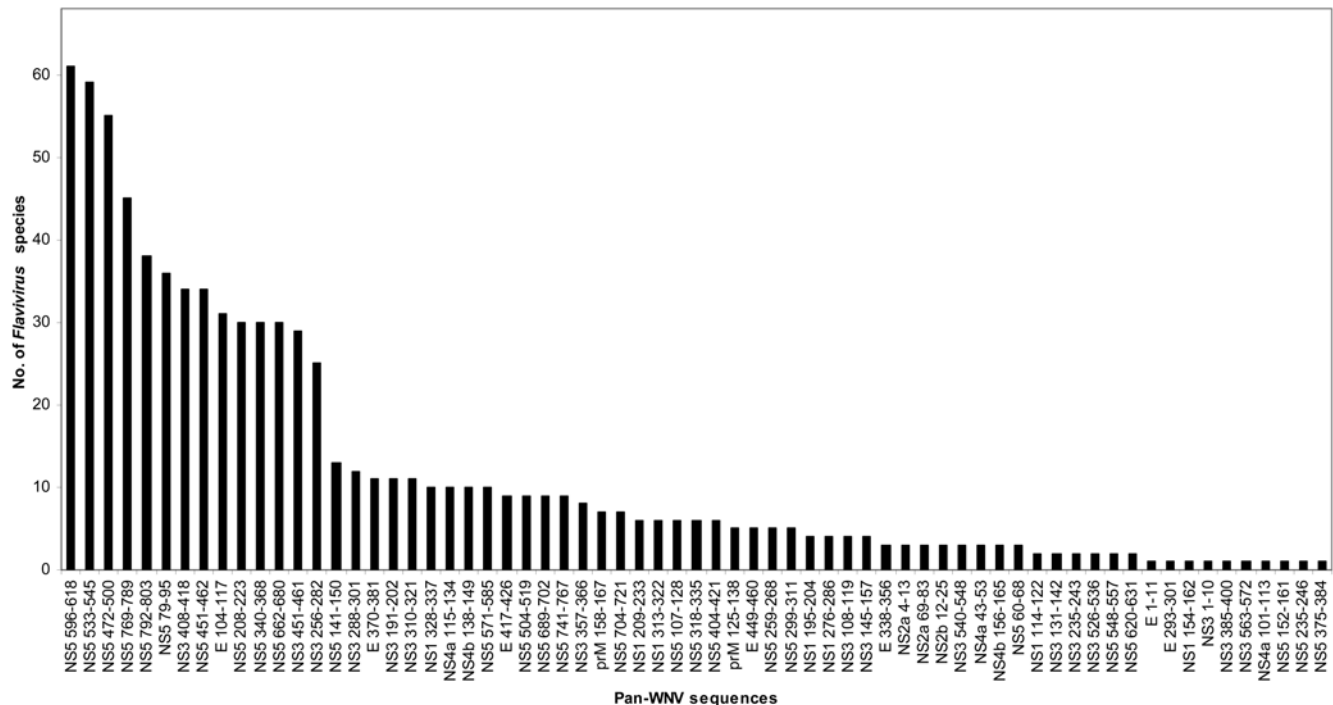


Figure 5. Number of flaviviruses shared by the pan-WNV sequences.
doi:10.1371/journal.pone.0005352.g005

previously reported WNV T-cell epitopes immunogenic in human, having HLA restriction (when known), with both class I (B*07) and II (DR2) specificities (Table 5). Further evaluation of the immune-relevance of pan-WNV sequences included a search for 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%) were predicted to contain 271 supertype-restricted binding nonamers (Figure 6 and Table S2). Of these sequences, 62 contained 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 [33,44], were found in 41 of the 78 sequences. Seven (7) of the 78 sequences had at least 3 sequential nonamers overlapping by 8 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 9 amino acids present in 54 CD4⁺CD8⁻ and/or CD4⁺CD8⁺ IFN- γ ELISpot positive peptides (Table 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 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 and Table S2). 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 9 of these 11 pan-WNV sequences were correctly predicted (Table S2). In summary, combined with previously reported data for human WNV T-cell epitopes from literature and public database (Table 5), at least 44 of the 88 pan-WNV sequences contained numerous HLA-restricted class I and/or class II epitopes demonstrated by *in vivo* T-cell response assays.

Discussion

In the 70 years following the discovery of WNV in Africa in 1937 [45], there has been 100% conservation of 88 pan-WNV sequences, 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, 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 [37,38], NS3 peptidase S7, DEAD/H domain [40,41], and NS5 proteins

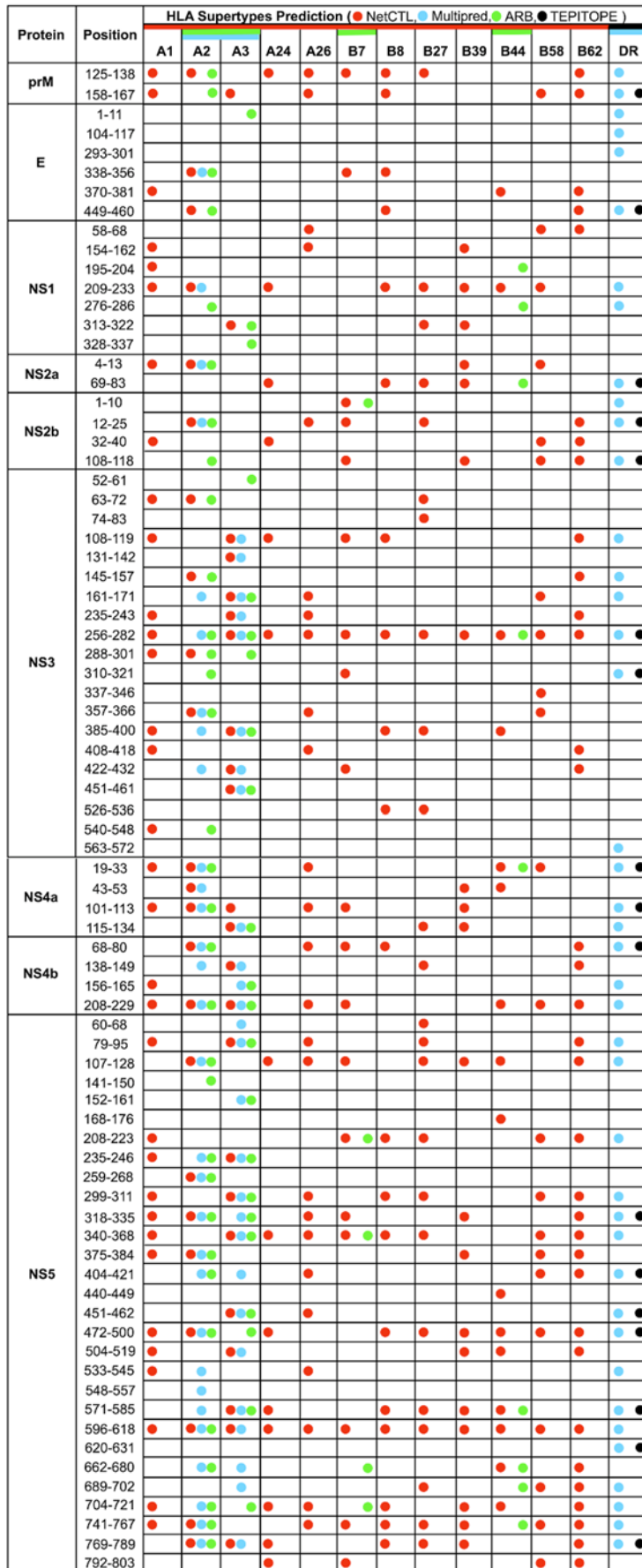


Figure 6. Putative HLA supertype-restricted, pan-WNV T-cell epitopes predicted by computational algorithms.
 doi:10.1371/journal.pone.0005352.g006

Table 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	<u>VIGLYGNGV</u>	CD4	DR2	[50]
	256-EIVDVMCHATLTHRLMSPHRVPNYNLF-282	<u>SPHRVPNYNL</u>	CD8	B07	[51]
NS5	704-GWYDWQVQPFCSNHFTL-721	<u>FCSNHFTL</u>	-	-	1021472

^aEpitope amino acids matching the pan-WNV sequences are underlined.

^b1021472 is an accession number of a record in the Immune Epitope Database.
doi:10.1371/journal.pone.0005352.t005

RdRp domain [42,43]. 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 9 consecutive amino acids of 5 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 [46,47]. Similarly, the NS3 nonamer sequence fragment found in the Asian Tiger mosquito (*Aedes albopictus*), is possibly due to genetic recombination between phyla [48]. 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 [49].

There is evidence that many of the conserved sequences are immunologically relevant in humans. Numerous (44/88) contained at least 9 amino acids overlapping with a total of 54 peptides that have been reported to be immunogenic in humans and/or HLA Tg 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 [17]. 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 [17].

A comparison can be made with a similar study of each of the four DENV serotypes [17]. In contrast to WNV, the sequences of the combined serotypes of DENV are highly diverse, 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; only two of these 44 sequences were completely conserved in all the four serotypes. However, the conservation and variability of each DENV serotype is comparable to WNV. The individual DENV serotypes and the WNV show remarkable stability over the entire recorded history of their sequences, as demonstrated by their low peptide entropies and variant frequencies. The pan-WNV sequences matched sequences of

DENV (one or more serotypes) with representations ranging from low (3%) to high (100%); similar observations were made for pan-DENV sequences matching WNV sequences (4 to 100%). The conserved sequences that matched with low representation may pose potential risk of altered ligands resulting in pathologic immune responses following co-infection or vaccination and secondary infection with a similar virus. Thus, while the consequences of such extensive possible cross-reactive immunity are hypothetical, we propose, for vaccine formulation, that it is prudent to select conserved sequences that are specific to the pathogen, and thus representative of a minimal number of variant sequences.

Supporting Information

Figure S1 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.

Found at: doi:10.1371/journal.pone.0005352.s001 (12.01 MB DOC)

Table S1 Percentage representation of pan-WNV sequences in other flaviviruses.

Found at: doi:10.1371/journal.pone.0005352.s002 (0.12 MB DOC)

Table S2 Putative HLA supertype-restricted binding nonamer peptides in pan-WNV sequences, predicted by immunoinformatics algorithms (NetCTL, Multipred (MP), ARB and TEPITOPE (TP)).

Found at: doi:10.1371/journal.pone.0005352.s003 (0.79 MB DOC)

Data Set S1 GI numbers of West Nile virus proteins downloaded from the NCBI Entrez Protein Database.

Found at: doi:10.1371/journal.pone.0005352.s004 (0.12 MB XLS)

Acknowledgments

The authors thank Paul Nordstrom August for his help with the illustrations, and Heiny Tan, Hu Yongli and George, Au Yeung Kiu Chi for their valuable suggestions.

Author Contributions

Conceived and designed the experiments: AMK JTA. Performed the experiments: QYK. Analyzed the data: QYK AMK KOJ SR TWT VB JS JTA. Contributed reagents/materials/analysis tools: KOJ OM JTA. Wrote the paper: QYK AMK JTA.

Table 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 [*]	<u>LVKTESWILRNPGYALVA</u>	DR2 & DR4
		<u>LRNPGYALVAAVIGWML</u>	A24, B7 & DR2
	158-LLLLVAPAYS-167 [*]	<u>RVVFFVLLLLVAPAYS</u>	A2, DR2, DR3 & DR4
E	104-GCGLFGKGSIDTCA-117 [*]	<u>RGWGNCGLFGKGS</u>	DR3 & DR4
	293-LKGTTYGVC-301 [*]	<u>EKLQLKGTTYGVC</u>	DR4
	370-ELEPPFGDSYIV-381	<u>KVLIELEPPFGDSYIV</u>	DR4
	449-LFGGMSWITQGL-460 [*]	<u>FRSLFGGMSWITQGL</u>	A2, DR2 & DR3
NS1	209-TWKLRAVLGEVKSCTWPETHLWG-233 [*]	<u>RLNDTWKLRAVLGEVK</u>	DR4
	276-DFDYCPGTTVT-286 [*]	<u>EGRVEIDFDYCPGTTVT</u>	DR4
	313-CRSCTLPPLR-322	<u>GKLITDWCCRSCTLPPLR</u>	DR3 & DR4
	328-GCWYGMIEIRP-337	<u>SGCWYGMIEIRPQRHDEK</u>	DR4
NS2a	69-NSGGDVVHLALMATF-83 [*]	<u>FAESNSGGDVVHLALMA</u>	DR4
NS2b	1-GWPATEVMTA-10 [*]	<u>GWPATEVMTAVGLMFAIV</u>	DR4
	108-SAYTPWAILPS-118 [*]	<u>ISAYTPWAILPSVVGFWI</u>	A24, B7 & DR4
NS3	1-GGVLWDTPSP-10	<u>GGVLWDTPSPKEYKK</u>	B7 & DR4
	52-TTKGAALMSG-61	<u>WHTTKGAALMSGEGRL</u>	DR3
	074-EDRLCYGGPW-083	<u>GSVKEDRLCYGGPWKLQH</u>	A2
	145-DVIGLYGNGVIMP-157	<u>PIVDKNGDVIGLYGNGVI</u>	A2
		<u>VIGLYGNGVIMPNGSYI</u>	A2
		<u>YISAIQGERMDEPIPA</u>	A2 & DR2
		<u>MLRKKQITVLDLHPGAGK</u>	A2 & DR2
	235-ALRGLPIRY-243	<u>VLDLHPGAGKTRRILPQI</u>	DR2
		<u>VAAEMAEALRGLPIRY</u>	DR4
		<u>EALRGLPIRYQTSVAVPR</u>	DR4
		<u>PREHNGNEIVDMCHATL</u>	A2, DR2 & DR4
	256-EIVDMCHATLTHRLMSPHRVNPYNLF-282 [*]	<u>IVDMCHATLTHRLMSPH</u>	DR2
		<u>TLTHRLMSPHRVNPYNLF</u>	A2 & DR2
		<u>KVELGEEAAIFMTATPPG</u>	A2
		<u>LQTEIPDRAWNSGYEWI</u>	A2
	310-AAIFMTATPPG-321	<u>EMGANFKASRVIDSRKSV</u>	A2
	337-QTEIPDRAWN-346	<u>CYGGHTNEDDSNFAHW</u>	A2 & DR3
	422-RVIDSRKSVKP-432	<u>AHWTEARIMLDNINM</u>	A2 & DR3
	470-GDEYCYGGHTNEDDSN-485	<u>EYRLRGEERKNFLELLR</u>	A2 & DR2
	487-AHWTEARIM-495	<u>DRRWCFDGPRTNTIL</u>	DR3
526-LRGEERKNFLE-536	<u>HFMGKTWEALDTMYVVA</u>	DR2 & DR4	
563-WCFDGPRTNT-572 [*]	<u>VLIPEPEKQRSQTDNQLA</u>	DR4	
NS4a	19-KTWEALDTMYVVA-33 [*]	<u>GEFLDLRPAWWSLYAV</u>	DR2
NS4b	115-MIVLIPEPEKQRSQTDNQLA-134 [*]	<u>PATAWSLYAVTTAVLTPL</u>	DR2 & DR3
	39-PATAWSLYA-47	<u>DYINTSLTSINVAQASALF</u>	DR3 & DR4
	68-TSLTSINVAQASAL-80 [*]	<u>LITAAAVTLWENGASSVW</u>	DR3 & DR4
	208-VTLWENGASSVWNTAIGLCH-229 [*]	<u>LVQSYGWNIVTMKSGVDV</u>	DR3
NS5	107-GPGHEEPQLVQSYGWNIVTMKS-128 [*]	<u>CDIGESSSAEVEEHRTI</u>	B7
	152-SSAEVEEHRT-161	<u>SAEVEEHRTIRVLEMV</u>	A2, B7 & DR2
	208-RNPLSRNSTHEMYWVS-223 [*]	<u>SRNSTHEMYWVSASGNV</u>	DR2
	451-TCIYNMMGKREK-462	<u>ECHTCIYNMMGKREKK</u>	A2
	472-GSRAIWFMWLGARFLEFALGFLNEDHWL-500	<u>AKGSRAIWFMWLGARFL</u>	A24
		<u>WFMWLGARFLEFALGFL</u>	A24
	596-ISREDQRGSGQVVYALNTFTNL-618 [*]	<u>REDQRGSGQVVYALNTF</u>	DR2

Table 6. cont.

WNV protein	Pan-WNV sequence ^a	T-cell epitopes immunogenic in HLA transgenic mice	
		ELISpot activation peptide ^b	ELISpot positive HLA transgenic mouse
		<u>GQVVTYALNTFTNLAVQL</u>	DR2 & DR4
	620-VQLVRMMEGEGV-631*	<u>NTFTNLAVQLVRMMEGEGV</u>	DR4
	704-GWYDWQQVPFCSNHFTL-721*	<u>GWYDWQQVPFCSNHFTL</u>	DR4
	741-GRARISPGAGWNVVRDTACLAKSYAQM-767	<u>DTACLAKSYAQMWLLLYF</u>	A24
	769-LLYFHRRLRLMANAICSAVP-789*	<u>YAQMWLLLYFHRRLRLM</u>	B7 & DR4
	792-WVPTGRTTWSIH-803	<u>NWVPTGRTTWSIHAGGEW</u>	DR4

^aPan-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 (*).

^bEpitope amino acids matching the pan-WNV sequences are underlined.

doi:10.1371/journal.pone.0005352.t006

References

- Horga MA, Fine A (2001) West Nile virus. *Pediatr Infect Dis J* 20: 801–802.
- Petersen LR, Roehrig JT (2001) West Nile virus: a reemerging global pathogen. *Emerg Infect Dis* 7: 611–614.
- Marfin AA, Petersen LR, Eidson M, Miller J, Hadler J, et al. (2001) Widespread West Nile virus activity, eastern United States, 2000. *Emerg Infect Dis* 7: 730–735.
- Lanciotti RS, Roehrig JT, Deubel V, Smith J, Parker M, et al. (1999) Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* 286: 2333–2337.
- Petersen LR, Marfin AA, Gubler DJ (2003) West Nile virus. *JAMA* 290: 524–528.
- Hayes EB, Sejvar JJ, Zaki SR, Lanciotti RS, Bode AV, et al. (2005) Virology, pathology, and clinical manifestations of West Nile virus disease. *Emerg Infect Dis* 11: 1174–1179.
- Domingo E (1997) RNA virus evolution, population dynamics, and nutritional status. *Biol Trace Elem Res* 56: 23–30.
- Elena SF, Carrasco P, Daros JA, Sanjuan R (2006) Mechanisms of genetic robustness in RNA viruses. *EMBO Rep* 7: 168–173.
- Bondre VP, Jadhav RS, Mishra AC, Yergolkar PN, Arankalle VA (2007) West Nile virus isolates from India: evidence for a distinct genetic lineage. *J Gen Virol* 88: 875–884.
- Lanteri MC, Heitman JW, Owen RE, Busch T, Geffer N, et al. (2008) Comprehensive analysis of west nile virus-specific T cell responses in humans. *J Infect Dis* 197: 1296–1306.
- Wang Y, Lobigs M, Lee E, Mullbacher A (2003) CD8+ T cells mediate recovery and immunopathology in West Nile virus encephalitis. *J Virol* 77: 13323–13334.
- Cheng Y, King NJ, Kesson AM (2004) Major histocompatibility complex class I (MHC-I) induction by West Nile virus: involvement of 2 signaling pathways in MHC-I up-regulation. *J Infect Dis* 189: 658–668.
- Shrestha B, Wang T, Samuel MA, Whitby K, Craft J, et al. (2006) Gamma interferon plays a crucial early antiviral role in protection against West Nile virus infection. *J Virol* 80: 5338–5348.
- Brien JD, Uhrlaub JL, Nikolich-Zugich J (2007) Protective capacity and epitope specificity of CD8(+) T cells responding to lethal West Nile virus infection. *Eur J Immunol* 37: 1855–1863.
- Purtha WE, Myers N, Mitaksov V, Sitati E, Connolly J, et al. (2007) Antigen-specific cytotoxic T lymphocytes protect against lethal West Nile virus encephalitis. *Eur J Immunol* 37: 1845–1854.
- Khan AM, Miotto O, Heiny AT, Salmon J, Srinivasan KN, et al. (2006) A systematic bioinformatics approach for selection of epitope-based vaccine targets. *Cell Immunol* 244: 141–147.
- Khan AM, Miotto O, Nascimento EJ, Srinivasan KN, Heiny AT, et al. (2008) Conservation and variability of dengue virus proteins: implications for vaccine design. *PLoS Negl Trop Dis* 2: e272.
- Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, et al. (2007) Evolutionarily conserved protein sequences of influenza A viruses, avian and human, as vaccine targets. *PLoS ONE* 2: e1190.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 33: D39–45.
- McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32: W20–25.
- O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, et al. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 3: 275–284.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.
- Thompson JD, Higgins DG, Gibson TJ (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: 4673–4680.
- Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 623–656.
- Miotto O, Heiny A, Tan TW, August JT, Brusica 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.
- Rammensee HG (1995) Chemistry of peptides associated with MHC class I and class II molecules. *Curr Opin Immunol* 7: 85–96.
- Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. *Nucleic Acids Res* 34: D227–230.
- de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, et al. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 34: W362–365.
- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. *Nucleic Acids Res* 32: D138–141.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242.
- Peters B, Sidney J, Bourne P, Bui HH, Buus S, et al. (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 3: e91.
- Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, et al. (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: 2295–2303.
- Zhang GL, Khan AM, Srinivasan KN, August JT, Brusica V (2005) MULTI-PRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res* 33: W172–179.
- Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics* 57: 304–314.
- Bian H, Hammer J (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. *Methods* 34: 468–475.
- Valdar WS (2002) Scoring residue conservation. *Proteins* 48: 227–241.
- Kanai R, Kar K, Anthony K, Gould LH, Ledizet M, et al. (2006) Crystal structure of west nile virus envelope glycoprotein reveals viral surface epitopes. *J Virol* 80: 11000–11008.
- Mukhopadhyay S, Kim BS, Chipman PR, Rossmann MG, Kuhn RJ (2003) Structure of West Nile virus. *Science* 302: 248.
- Li H, Clum S, You S, Ebner KE, 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: 3108–3116.
- Erbel P, Schiering N, D'Arcy A, Renatus M, Kroemer M, et al. (2006) Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. *Nat Struct Mol Biol* 13: 372–373.
- Feito MJ, Gomez-Gutierrez J, Ayora S, Alonso JC, Peterson D, et al. (2008) Insights into the oligomerization state-helicase activity relationship of West Nile virus NS3 NTPase/helicase. *Virus Res* 135: 166–174.
- Mackenzie JM, Kenney MT, Westaway EG (2007) West Nile virus strain Kunjin NS5 polymerase is a phosphoprotein localized at the cytoplasmic site of viral RNA synthesis. *J Gen Virol* 88: 1163–1168.
- Malet H, Egloff MP, Selisko B, Butcher RE, Wright PJ, et al. (2007) Crystal structure of the RNA polymerase domain of the West Nile virus non-structural protein 5. *J Biol Chem* 282: 10678–10689.

44. Zhang GL, Khan AM, Srinivasan KN, Heiny A, Lee K, et al. (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.
45. Smithburn KC, Hughes TP, Burke AW, Paul JH (1940) A neurotropic virus isolated from the blood of a native of Uganda. *American Journal of Tropical Medicine* 20: 471–492.
46. Biswas T, Aihara H, Radman-Livaja M, Filman D, Landy A, et al. (2005) A structural basis for allosteric control of DNA recombination by lambda integrase. *Nature* 435: 1059–1066.
47. Gottesman ME, Weisberg RA (2004) Little lambda, who made thee? *Microbiol Mol Biol Rev* 68: 796–813.
48. Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, et al. (2004) Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. *J Gen Virol* 85: 1971–1980.
49. Freeman S (2003) *Biological Science*. NJ, USA: Prentice Hall.
50. Kurane I, Okamoto Y, Dai LC, Zeng LL, Brinton MA, et al. (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): 2243–2249.
51. De Groot AS, Saint-Aubin C, Bosma A, Sbai H, Rayner J, et al. (2001) Rapid determination of HLA B*07 ligands from the West Nile virus NY99 genome. *Emerg Infect Dis* 7: 706–713.