INSTITUTO POLITÉCNICO NACIONAL

Centro de Biotecnología Genómica

Genome sequencing of dengue virus isolated from Mexico and its large-scale sequence comparative analysis

T E S I S

Que para obtener el grado de:

Doctor en Ciencias en Biotecnología

Presenta:

Edgar Eduardo Lara Ramírez

Cd. Reynosa, Tamaulipas, México, Noviembre de 2014.

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de Reynosa, Tamps, el día 7 del mes Noviembre del año 2014, el que suscribe Edgar Eduardo Lara Ramírez alumno del Programa de Doctorado en Ciencias en Biotecnología con número de registro B101593 adscrito a Centro de Biotecnología Genómica manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección de Dr. Xianwu Guo y cede los derechos del trabajo intitulado "Genome sequencing of dengue virus isolates from Mexico and its large-scale sequence comparative analysis" al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección Blvd. del Maestro esq. con Elías Piña S/N Col. Narciso Mendoza, C.P. 88710 Cd. Reynosa, Tamaulipas, México Tels. 01-899 9243627, 9251656. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Edgar Eduardo Lara Ramírez

Nombre y firma

SIP-14-BIS

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de Reynosa, Tamps. siendo las 13:00 horas del día 07 del mes de Noviembre del 2014 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación de CBG para examinar la tesis titulada:

"Genome sequencing of dengue virus isolates from Mexico and its large-scale sequence comparative analysis"

Presentada por el alumno:

Lara	Ramírez	Edgar Eduardo
Apellido paterno	Apellido materno	Nombre(s)

Con registro: B 1 0 1 5 9 3

aspirante de:

Doctorado en Ciencias en Biotecnología

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de tesis

Dr. Xianwu Guo

Dr. Juan Santiago Salas Benito

Dr. Mario Alberto Rodríguez Pérez

Dr. José Alberto Narváez Zapata

Dra. Ana María Sifuentes Rincón

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Miguel Ángel Reyes López

INSTITUTO POLITECNICO NACIONAL
CENTRO DE BIOTECNOLOGIA GENOMICA

This work was carried out at the Laboratory of Molecular Biomedicine, Center for Genomic Biotechnology, National Polytechnic Institute, under the direction of Dr. Xianwu Guo and Dr. Juan Santiago Salas Benito.

Dedicated to my beloved daughter Dulce and my wife Sori

Indice of contents

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

To the National Polytechnic Institute, a leader in cutting-edge technology education in Mexico, for giving me the necessary tools in my academic development.

To the Center for Genomic Biotechnology for the facilities to perform the necessary activities to accomplish my Doctor in Science program.

To the CONACyT for support me with the scholarship.

To my supervisor and mentor Dr. Xianwu Guo, for guide me during this research project.

To all the people involved in this life project

ABREVIATIONS

Å. Armstrong.

DENV. Dengue virus

CNS. Central Nervous System

DF. Dengue Fever

DHF. Dengue Hemorrhagic Fever

DSS. Dengue Shock Syndrome

IgM. Immunoglobulin M

μL. Microliters

GC. Guanines and Citocines

DNA. Deoxyribonucleic Acid

RNA. Ribonucleic Acid

%. Percent

nm. Nanometers

h. hora

°C. Grades centigrade

min. minute or minutes

PCR. Polymerase Chain Reaction

rpm. Revolutions per minute

ng. nano grames

IEDB. Immune Epitope DataBase

ER. Endoplasmic Reticulum

RSCU. Relative Synonimous Codon Usage

CA. Correspondence Analysis

NIH. National Institutes of Health

MHC. Major Histocompatibility Complex

RESUMEN

El dengue representa un importante problema de salud pública en América debido a que la incidencia de esta enfermedad se ha incrementado notablemente en los últimos años, cambiado regiones hipoendémicas a regiones hiperendémicas. El análisis filogenético del gen E en cepas del virus del dengue (DENV; serotipos 1-4) de México mostró la ausencia de introducción de nuevos serotipos en el país hasta el 2010. El análisis de genomas a gran escala enfocado en el uso de codones y usando el análisis de correspondencias reveló que cada una de las cepas de DENV1-4, incluyendo las de México, se encuentra influenciada por su origen geográfico. En un análisis de correlación en un contexto global del eje con mayor fuente de variación en los genomas contra el contenido de GC en las tres posiciones de nucleótidos (codón de GC1, GC2 y GC3), así como el número efectivo de codones indicó que la presión de mutación es uno de los principales factores que influyen en el uso de codones, pero con distinto nivel de acuerdo a la posición específica de nucleótidos en el codón. Por otra parte, nuestro estudio mostró que no sólo el GC3, sino también el GC1 y el GC2 tienen una buena correlación con el eje de mayor variación, lo que sugiere que todos los sitios de codones están relacionados con la agrupación de las cepas geográficas, incluidas las cepas mexicanas. El análisis detallado de las firmas genómicas de aminoácidos en la proteína de la envoltura de DENV-1-4 mostraron que el patrón de aminoácidos de las cepas mexicanas es muy similar a las cepas de América del Norte y América del Sur. Además, el análisis estadístico multivariado confirmó la alta influencia del origen geográfico. El análisis de epítopes permitió identificar aquellos que podrían estar relacionados con el desarrollo de la enfermedad. El presente análisis en un contexto global de secuencias de DENV, incluyendo secuencias mexicanas, revela información valiosa sobre la evolución genómica, epidemiología e inmunología de los DENV.

ABSTRACT

Dengue represents a major public health problem in the Americas because the dengue incidence has been markedly increased in recent years so that the Americas have been changed from hypoendemic regions to hyperendemic regions. The E gene phylogenetic analysis for Mexican DENV 1-4 sequences showed the absence of new serotype strains in the country until 2010. The large-scale codon usage genome analysis using correspondence analysis revealed that the DENV1-4 Mexican strains are located in the North and South American clusters. This finding showed that Mexican strains, along with the other American strains, have a limited codon usage restrained by their geographic origin. In a global context, the correlation of the first axis position in the correspondence analysis of the genomes with the GC content at the three nucleotide positions of codon (GC1, GC2 and GC3) as well as the Effective Number of Codons revealed that the mutation is one of the major forces influencing the codon usage, but with distinct pressure on specific nucleotide position in the codon. Furthermore, our study showed that not only GC3, but also GC1 and GC2 have a good correlation with Axis major variation, suggesting that all the codon sites are related to clustering of geographical strains, including the Mexican strains. The detailed analysis on the genomic amino acid signatures on the envelope protein of DENV-1-4 showed that the amino acid pattern of the Mexican strains is quite similar to the North American and South American strains. Further, the multivariate statistical analysis confirmed the strong influence of the geographic origin on aminoacid frequency. Some epitopes were identified to be related to disease development. These analyses in global DENV sequences along with Mexican sequences provided new information on DENV genomic evolution, epidemiology and immunology.

1. INTRODUCTION

Dengue viruses (DENV) are RNA viruses of the family of *Flaviridae* and they are causative agents of dengue fever (DF), the most prevalent arthropod-born viral disease worldwide, and they are considered as an emerging global health threat. There are four related but antigenically distinct DENV serotypes (Beaumier and Rothman, 2009).

The principal vectors of DENV are mosquitoes of *Aedes* genus, predominately *Aedes aegypti*. DENV are classified as re-emerging pathogens by National Institute of Allergy and Infectious Disease (NIAID), an organization of the United States National Institutes of Health (NIH). These pathogens are now causing disease in areas where human was heretofore not infected, or are causing more severe diseases in areas where only mild disease once occurred (Vaughan et al., 2010)

Genome sequence is basic information, which could be used to analyze the evolution and biological function of an organism by of comparative genomics techniques. The increasing information of molecular data accumulated in public databases about dengue genomes offers an opportunity to study the genomic nature of DENV.

The main objective of this study was to perform a large-scale sequence analysis through comparative genomics in order to understand the evolutionary molecular mechanisms and adaptation to the host for DENV 1-4, including the available Mexican sequences. The structure of the thesis comprehends four objectives in order to understand some evolutionary aspects of DENV, which include Mexican strains.

2. BACKGROUND

2.1. Dengue disease

The causal agent of dengue disease is dengue virus (DENV), which is transmitted to humans by the bite of mosquitoes of the genus *Aedes*. After the DENV infection is achieved, two distinct clinical syndromes may develop. These two syndromes are dengue fever (DF) and dengue hemorrhagic fever (DHF) (WHO, 2011). While these two clinical entities can be considered as a result of the same pathological process and the same viral agent could be implicated, the signs and symptoms differ from each other. These syndromic differences can be explained by the kind of host immune response, the genetic load from the host and the kind of DENV serotype.

2.1.1. Dengue fever

DF is the most common clinical form of DENV infection; it begins after an average incubation period of 2-8 days and is distinguished by three clinical stages: i) Prodromal phase: This quiescent phase is characterized by symptoms such as nasal congestion, rhinorrhea, sore throat and tearing, occasionally with conjunctivitis. At this phase the fever is absent. The half of clinical cases occurs only in adults, who are more prone to this symptomatology (Jelinek, 2000). ii) Clinical Phase: This phase is characterized by high fever with chills, frontal headache, myalgias, arthralgias, and retrorbital pain. The fever lasts for three to five days, and has a biphasic intermittent pattern, when it remits may cause rash, which indicates the disappearance of the virus from the blood stream (Gurugama et al., 2010). This phase occurs in both the adult and the pediatric population and (Guha-Sapir and Schimmer, 2005). iii) Convalescence phase. This stage only occurs in the adult population in a small number of

cases; it is characterized by depression, fatigue and weakness for a prolonged period (Murillo-Llanes et al., 2007).

DF is considered a benign pathogenic process and it does not require great efforts for treatment handling. However, the clinical features of this disease may mimic other viral infections. Thus is important to be familiar with the clinical presentation to improve the diagnosis and monitoring to establish their possible evolution to DHF (Jelinek, 2000).

2.1.2. Dengue hemorrhagic fever

DHF may develop two or three days after DF is diagnosed. This severe DENV infection is manifested by a high vascular permeability and hemostatic changes, related to the presence of haemoconcentration due to the plasma leakage to the extravascular space (Avirutnan et al., 2006). The signs and symptoms are characterized frequently by the presence of bleeding (epistaxis, bleeding gums, urogenital bleeding, bleeding from puncture sites, hemoptysis and bleeding of the digestive tract) and extravasation of fluids (bruising or petechiae). The laboratory tests often have shown a hematocrit of less of 20% or thrombocytopenia less than 100,000 platelets per cubic milliliter (Murillo-Llanes et al., 2007). DHF can evolve to more severe clinical form called dengue shock syndrome (DSS), that cause a systemic circulatory failure manifested by a rapid and weak pulse, hypotension and it has high mortality (WHO, 2011).

2.1.3. Epidemiological evolution of dengue

The first illness clinically related to DF described as break-bone fever was documented by Benjamin Rush in Philadelphia in 1780 (Rush, 1789). Nevertheless, dengue disease was recognized at the end of 18th century in French West India, Batavia, Cairo and Philadelphia (Carey, 1971).

The expansion of dengue around the world began near to the end of World War II as a result of the large-scale urbanization processes due to the increase of birth rate and life expectancy which favored the colonization of previously man free areas by the mosquito vector (Murillo-Llanes et al., 2007). However, during the expansion of dengue disease, the serotypes were no dispersed simultaneously, thus the production of outbreaks and epidemics over time on the countries have been attributed to new serotype introductions or mutations (Carrillo-Valenzo et al., 2010).

The DHF is considered a contemporary disease of this century. The first case was reported in 1954 in the Philippines; onwards dengue disease has been well-known around the world. Thailand reported its first DHF cases in 1958, and in less than one year other regions, such as Asia and the Pacific reported similar epidemics (Montes, 2001). In the 1980s, DHF began a second expansion into Asia when Sri Lanka, India and the Maldives reported their first outbreaks and China reported cases of hemorrhagic dengue after 35 years of absence. In all the countries from Asia where DHF is considered endemic, the outbreaks have worsened over the past 15 years. In the Pacific, dengue viruses were reintroduced in the 70s, after more than 25 years without reporting confirmed cases (Montes, 2001). In Africa, FD epidemics caused by all serotypes have increased since the 1980s. In 1970, the serotype 2 predominates in the Americas. However DENV3 had a focal distribution in Colombia and Puerto Rico. In

1977, DENV1 was introduced, which resulted in outbreaks that persist in the region for a period of 16 years. In 1981, DENV4 was introduced, also causing epidemics, mainly in the Caribbean (Gomez-Dantes and Willoquet, 2009). In 1981, a strain of DENV2 from Southeast Asia led to the first major outbreak of DHF in the Americas. A strain, also of Asian origin, expanded quickly through the region and caused outbreaks of DHF, eight years after the first one of DHF in Venezuela, Colombia, Brazil, French Guyana, Suriname and Puerto Rico (Montes, 2001). In 1997, 18 countries of America reported confirmed cases of DHF, which is now endemic in many developing countries, including Mexico. The dramatic rise in the number of dengue cases over the last decade has been associated to demographic changes and ecological alterations due to global warming, which facilitates vector breeding and facilitates human contact with mosquitoes.

The World health organization estimates that 390 million infections occur annually, in over 100 countries (Figure 1); 96 million of those manifest clinically (Bhatt et al., 2013).

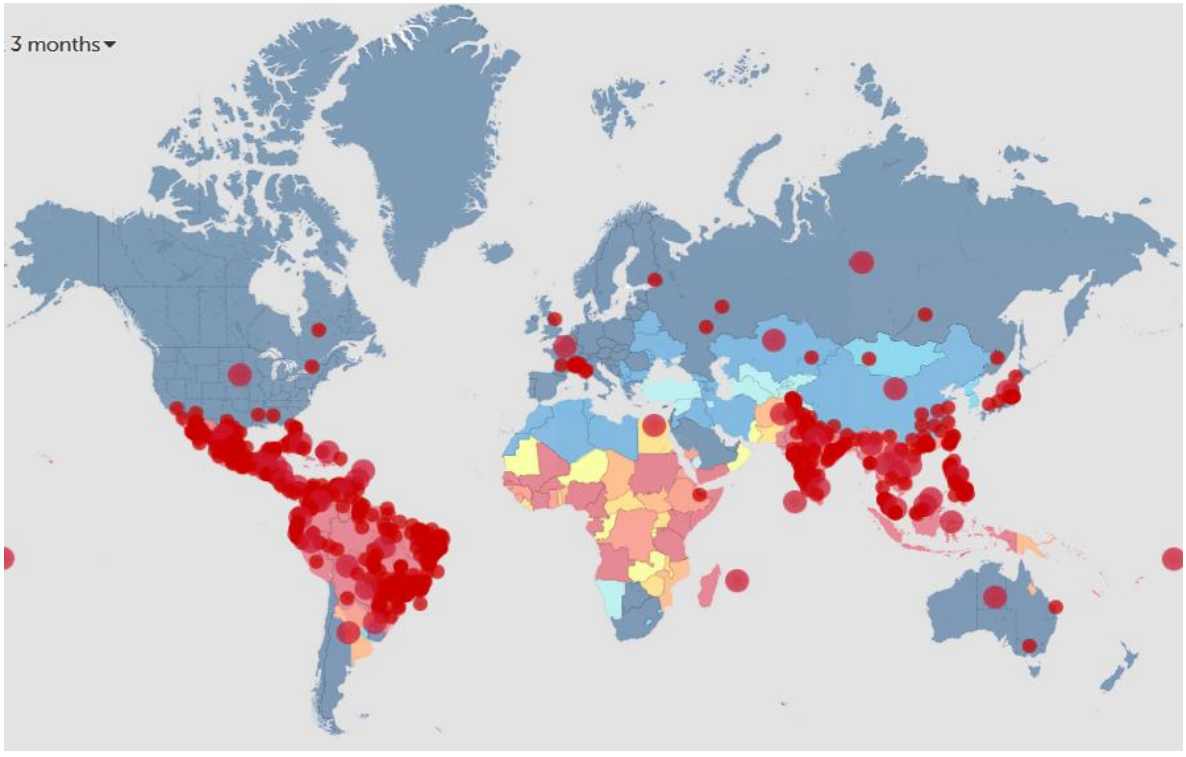


Figure 1. World distribution of dengue. From http://www.healthmap.org/dengue/es/

In Mexico, dengue is also a major public health problem. With the control of *A. aegypti* in the decade of the 60s, Mexico was free of dengue until 1978 when it was re-invaded again by the primary vector. Since then, the cases attributed to DENV have an annual pattern with peaks in the rainy months. Currently the four DENV serotypes are circulating in México (Carrillo-Valenzo et al., 2010).

2.1.4. Diagnosis

After the DF or DHF is diagnosed a confirmatory test is obligatory. The confirmation of dengue is done through laboratory techniques (Alexander Diaz-Quijano et al., 2006), that include:

Viral isolation.- it can be carried out in two forms: in the first, a presumable sample infected by DENV is inoculated on newborn intracerebrally to induce the developing of CNS disease with distinct clinical features. This form is less sensitive and is not recommended at present. The most used form is the viral isolation from cell lines from mammalian or mosquito cells, being the last ones the highest sensitivity for diagnosis (Guzmán and Vázquez, 2002).

Viral identification.- This test is a fast, simple and inexpensive method for the diagnosis, it can be done through monoclonal antibodies that generate hybridomas independently from the cell culture source. It requires a large amount of samples due to the poor viral replication obtained using antibodies (Acosta-Bas and Gómez-Cordero, 2005).

Antigenic test.- It consists in the antigen identification on the patient serum. It is a rapid and specific technique. The disadvantage is the low sensitiveness and it does not determine the kind of DENV serotype (Guzmán and Vázquez, 2002).

Serologic test.- This is the most used and fast method employed in epidemic situations; it consists in the identification of antibodies in the serum of patients for detecting both acute disease (IgM detection, that appears from the fifth day of infection and remains positive up to 90 days of after withdrawal of the disease) and memory (IgG during the rest of life) (Alexander Diaz-Quijano et al., 2006). These tests do not discern between each of the four DENV serotypes. However, the study by Delgado et al., (Delgado et al., 2002) reports that is possibly the serotype identification through the titer levels of IgM, but only in countries where two serotypes are circulating.

Molecular test.- The molecular techniques is the most reliable way to establish the diagnosis in terms of serotype and is used in epidemiological surveillance when epidemics or outbreaks that may involve atypical mutations identified any serotype or geographical variant of any of them. The RT-PCR has become the best way to diagnose cases of dengue, which not only confirms the serotype and variant of this but it is also fast and successful using several kind of samples (serum, tissue, cell supernatants, insects, etc.) (Lanciotti et al., 1992).

2.1.5. Treatment

To establish a proper management for DENV infected people is necessary its segregation from uninfected people. This epidemiological measure allows take the control of the people that need treatment in order to save time and resources (WHO, 2011). Actually

symptomatic treatment is provided for DF, and for the severe forms, intensive fluid replacement could be necessary. At present there is no drug or vaccines to treat DENV infection (Rajapakse et al., 2014).

2.2. Dengue virus (DENV)

DENV is an arbovirus of the genus *Flavivirus* genus of the *Flaviviridae* family, which causes millions of infections worldwide and although its mortality rate is very low, the high number of cases reported in recent times exceed the total cases by death endorsed to this pathogen, obeying the importance of the study of this kind of viruses (Gubler, 2002).

2.2.1. Classification

DENV is grouped into the *Flaviviridae* family, very related to the *Togaviridae* because they share similar viral structure (Velandia and Castellanos, 2011). The *Flaviviridae* family includes the genera Flavivirus, Pestivirus and Hepacivirus (Hepatitis C virus); the first are the causal agents for the known arthropod-borne diseases (Guha-Sapir and Schimmer, 2005).

The Flaviviruses are 40 to 60 nanometers in size, and they have a nucleocapsid containing genetic material (single strand RNA with positive polarity) surrounded by an envelope. Those viruses have a similar appearance on the microscope view (Thurner et al., 2004). This genus species is variously according to the vector transmitter and within these species, DENV are one of the most important viruses transmitted by mosquitoes. Within DENV species exists a sub classification based on immunological and molecular criteria that sorted DENV into four serotypes, such as: DENV-1, DENV-2, DENV-3 and DENV-4. Each

DENV1-4 creates lifetime immunity against re-infection of the same serotype (homologous) and a short cross-immunity against the other three serotypes (heterologous), which can last several months. The four strains are capable of producing DF as well as the severe forms of the disease (DHF, DSS) that can lead to death. Some DENV genetic variants within serotypes appear to be more virulent or have greater pathogenic potential for epidemics than others.

2.2.2. Viral structure

DENV, consists of an RNA genome conformed by a icosahedral nucleocapsid (size approx. of 30 nm diameter), surrounded by a lipid bilayer or envelope of 10 nm thickness where two viral proteins are associated originating the 50 nm diameter virion (Velandia and Castellanos, 2011). The lipid bilayer is associated with the E and M protein. The viral particles released into the extracellular medium are indistinguishable from those found in the intracellular vesicles (Modis et al., 2003). The immature particles contain only the unprocessed precursor of the protein prM, and are less pathogenic than the released particles (Tomlinson et al., 2009). In relation to the physical-chemical characteristics for DENV, they possess a density of 1.23 gr/cm^3 and a sedimentation coefficient of 210S (Henchal and Putnak, 1990).

The lipid composition of the virus is poorly understood, reflecting the composition of the host cell membrane, where the budding occurs and they are probably intracellular membranes of the ER. Total lipids are 7% of virion dry weight; of which 90% are phospholipids, 7% sphingomyelin, cholesterol and other neutral lipids (Henchal and Putnak, 1990). In the electronic microscope view we can identify these structures: The icosahedral virion composed asymmetrically by 90 subunits (250 Å), each of which can be can hold up to 4 protein subunits of E. A greater depth of between 185-220 Å, a structure was observed with

only 60% of the surface with respect to the above but the presence of glycoproteins and E protein is identified in the higher definition. With a range of 140-185 Å is distinctive the lipid membrane including internal and external layer and the presence of phosphates. At 105-135 Å the nucleocapsid is visible, which represents 50% of the size of the virus (Henchal and Putnak, 1990).

2.2.3. DENV life cycle

Once the mosquito bite, DENV is internalized within dentritic cells of the skin by binding to the host cell through the two different mechanisms (Clyde et al., 2006)

Host cellular Receptor mechanism.- This mechanism is used by the virus during the primary infection. The mechanism involves the domain 3 of the envelope glycoprotein, which binds to multiple receptors of the host permissive cells (monocytes, macrophages and dentritic cells) to start the process of entry and fusion (Rodenhuis-Zybert et al., 2010). This mechanism is responsible of the infection for almost every body cell lines (Chen et al., 1997).

Inmunocomplex mechanism.- This process occurs during secondary infections and is responsible for immune attack to lymphocytes, monocytes and macrophages producing the host immunosuppression. In this mechanism, the virus is anchored in an immunoglobulin, which makes the cellular interaction via the Fc portion of the IG (Montes, 2001).

Once performed the virus-cell binding, a conformational change occurs in the cellular cytoskeleton permitting the induction of the heat shock proteins that aid for the fixation of the virus to the cell membrane and the total fusion virus-host cell membrane (Reyes-Del Valle et al., 2005). In the cytoplasm, the virus escapes from the endosome due to the acidic pH, that

promotes a change in the shape of the domain 2 of the viral glycoprotein allowing it`s anchoring to the endosomal membrane and releasing the nucleocapsid into the cytoplasm and the envelope viral membrane (Velandia and Castellanos, 2011).

Viral RNA replication. Once nucleocapsid is free in the cytoplasm, the cellular ribosomes initiate the translation. The viral genome has a single open reading frame for translation initiation of the viral protein which is held in the endoplasmic reticulum (ER) (Clyde et al., 2006). This polyprotein undergoes proteolitically cleavage mainly by a cellular signalase and the viral serin protease NS·3/BS2B. Once the RNA the RNA-Dependent RNA Polymerase (RdRp), has been synthesized (NS5) and the partial circularization of viral RNA is achieved the viral replications begins, (Alvarez et al., 2005; Erbel et al., 2006).

Assembly, Maturation and Release of Virus. The replication process of DENV viral genome occurs with the aid of the NS proteins. The new viral RNA is packaged to form the nucleocapsid with the assistance of the C protein. The immature virions formed in the ER contain heterodimers of the pr and E proteins oriented into its lumen (Rodenhuis-Zybert et al., 2010). Those heterodimers associate into oligomeric trimmers to induce the curved surface lattice that guides the virion budding. From the ER the immature travel through the secretory pathway until de Golgi apparatus. In the trans-Golgi network the viral particles undergoes a structural reorganization of the glycoproteins prM/E enabling the cellular endoprotease furin the cleavage of the prM generating the M and “pr” peptides. These two peptides stabilize the E protein preventing premature conformational changes that would lead to the membrane fusion (Kuhn et al., 2002; Zhang et al., 2004). Upon dissociation of the pr peptide, the mature virions are ready to infect other cells from the host

2.2.4. DENV genome

DENV virions possess a positive-sense, single-stranded RNA genome. The viral genome is approximately 11 kb in length and consists in a single open reading frame encoding a polypeptide of 3300 amino acids which is pos-translationally cleaved to produce a three structural proteins [Capsid (C), pre-membrane/membrane (prM) protein and envelope (E) protein], seven non-structural proteins (NS-1, -2A, -2B, -3, -4A, -4B and -5) and two non-translated region (5'NTR, 3'NTR) (Chambers et al., 1990). The virion proteins all arise by proteolytic processing from a large poly protein precursor. The structural proteins are located at the 5 'portion and the remaining non-structural proteins are located in the terminal segment (Alvarez et al., 2005).

2.2.4.1. Structural Proteins

The C protein also called the core protein, is the first synthesized polypeptide, has a weight of 11 kD and is composed of four helices with different functions (Filomatori et al., 2006). i) the helix one is located on the N-terminal with a cytoplasmic orientation that allows the anchoring to a newly synthetized RNA through its basic amino acids, protecting it from degradation and helps for shaping the nucleocapsid through their associations in homodimers (Jessie et al., 2004). ii) the helix two is hydrophobic and it is involved in the formation of the ribo nucleoprotein and virion; has a form of "hinge", that facilitates the association of the viral RNA to the endoplasmic reticulum where nucleocapsid is produced. It is also allows small lipid droplets, where the C proteins is located, with the two other structural (E, M) proteins to complete the formation of the virus (Tomlinson et al., 2009). iii) the helixes three

and four allows the anchoring of the viral RNA to the ER through their hydrophobic properties.

The precursor membrane protein (prM) has a weight of 26 KD and is processed by furin protease in the late state of the replicative cycle, just before the virion release. This cleaveage generates a “pr” portion of 17 kD, that lead to the early formation of antibodies to this region, allowing the destruction of immature virions that still possess this portion (Chiu et al., 2005). The second section generates the M protein, which weighs only 8 kD and is present only in mature virions. The M protein is responsible for forming the membrane around the nucleocapsid because it has two transmembrane domains that allow the interaction with the protein C and the endoplasmic reticulum. Recent studies confirm the existence of an ectodomain 40 in the M protein, producing the so-called ApoptoM, which triggers cell apoptosis where the virus is conceived, allowing their release; thus the M protein through its ectodomain could be responsible for the tissue damage during infection (Khumthong et al., 2002).

Protein E is the envelope protein and the antigenic determinant of the virus. This protein is anchored to the viral membrane and is responsible for promoting interaction virus-cell host in the endocytosis process. It consists of three domains (I, II and III or A, B and C), that promote infection. The four serotypes of DENV that infect humans are distinguished by unique antigen determinants (epitopes) located on the envelope (E) protein. The E protein is the primary antigenic site for DENV and is responsible for inducing neutralizing antibodies and cell mediated immune responses in DENV-infected hosts.

2.2.4.2. No Structural Proteins

The NS1 protein has a weight of 46 kD and is not well known although their function has been determined as part of the viral assembly. This protein is highly antigenic since not only elicits humoral immunity but also complement activation. It can be found in the circulation and cytoplasm non-associated with the virion (Vaughn et al., 2000). This immunity however, generates damage because antibodies against it are able to recognize infected and healthy cells, including endothelial cells generating the plasma extravasation and dengue hemorrhagic fever (Ma et al., 2004)

The protein NS2A (22 kD) participates in the viral replication and assembly since it coordinates the generation of replicative or intermediates forms of the viral genome during the synthesis of RNA and it will associate the nucleocapsid to generate the mature virion (Velandia and Castellanos, 2011).

The NS2 protein of the DENV undergoes scission to form the portion corresponding to the NS2B, which has a weight of approximately 14 kD and is characterized by a hydrophobic region which serves to anchor together with the NS3 protein, to the membrane of the endoplasmic reticulum. Furthermore, the hydrophilic region allows activation of the protease activity of NS3 (D'Arcy et al., 2006) 3. The two roles that the NS2B in virus replication are: 1.-NS3-Stabiliser NS2B.-alone, the NS3 protein has no function because it requires that the chains be anchored by means of hydrogen bonds to the N-terminus of the protein NS2B, this portion allows the stabilization of the complex and remain attached to the endoplasmic reticulum, which it is a critical step (Erbel et al., 2006). 2.-activator NS2B.-NS3protein is in an inactive form until it is stabilized, so the protein of interest is essential for the initiation of viral replication. This activation is due to the binding of the portion C-

terminal and NTPase portion of NS2B and NS3 protease respectively (Khumthong et al., 2002).

The NS3 protein has a weight of 70 kD and is one of the most important for the replication process and perpetuity of this enzyme because it has three main functions (Erbel et al., 2006). NTPasa, this function is located at the C-terminal and is responsible for providing the energy needed for the replication process is carried out by inserting nucleotides to the positive strand of RNA. Helicase, this domain is located at the C-terminal and is responsible for unwinding the viral RNA in order that the RNA polymerase can bind on to start the process of viral translation. Protease effect, NS3pro domain is located at the N-terminal and acts as a protease to cleave some part of the viral polypeptide chain to generate mainly the non-structural viral proteins.

The NS4 protein is modified and generates two variants NS4A and NS4B, both are very hydrophobic and have a weight of 16 and 27 kD, respectively. The functions of these two proteins are a question, however it has been established that function as cofactors for replication (Khumthong et al., 2002). NS4B has been implicated in the evasion of interferon response

The NS5 protein is the most conserved among all flaviviruses. This protein is multifunctional, as the N-terminal end has enzymatic activity and guanidiltransferase methyltransferase is responsible for the capping and 5'methylation of the genomic RNA, while, at the C-terminal domain of RNA dependent RNA polymerase. Therefore, the NS5 protein acts as the only polymerase during replication and viral transcription (Tassaneetrithep et al., 2003). Although these processes take place entirely in the cytoplasm of infected cells,

has been in the nucleus; however, the reason and function of the NS5 in this compartments not completely understood (Zhang et al., 2003).

2.2.6. DENV Genome evolution

The organization and interrelation of the huge amount of heterogeneous information linked to genome analysis in order to find all the functional parts of genome sequences is one of the main focuses of genomics field (Hardison, 2003). The genomic information can helps in the analysis of phenotypic behaviors, designing experiments to know the relationship between structure and function or obtaining detailed understanding of molecular processes.

The genomic studies also provide clues about patterns of evolution from different perspectives; for example, can helps to uncover the heterogeneous nature of genes or genomics regions, or understand the factors that shape the evolutionary constraints of the lineages under study (Baldauf, 2003; Miller et al., 2004).

DENV has been also studied from the angle of viruses' evolution providing worth information. For example, several studies support that DENV has evolved from a nonhuman primate virus transmitted by mosquitoes before diverging into the four DENV serotypes (DENV-1 to DENV-4), that later emerged into the human population (Vasilakis et al., 2007b). Wang *et al*. (Wang et al., 2000) tested this hypothesis using phylogenetic analyses to compare envelope protein gene sequences of the endemic (human and *A. aegypti*) isolates to sequences of sylvatic (nonhuman primate and arboreal *Aedes spp*.) DENV-1, -2 and -4 strains of Southeast Asian origin, as well as DENV-2 sylvatic strains from West Africa. These analyses indicated that endemic DENV-1, -2 and -4 emerged independently from sylvatic progenitors

at a time, consistent with the establishment of urban populations in the Asia-Oceania (Kuno, 1995). The Asiatic origin of DENV sylvatic progenitors is supported by serological surveys of ecologically diverse rural habitats in Southeast Asia (Smith, 1956). Those evolutionary studies on dengue virus indicated the rise of distinct endemic genotypes within each serotype (Gubler, 1997).

Other evolutionary studies focused on selection pressure analysis and his correlation with disease considered that the most common pressure acting on DENV in nature is purifying selection, with little or no evidence of adaptive evolution (Zanotto et al., 1996). Contrariwise, a study reported adaptive evolution in DENV 3 and most strongly in DENV 4 (Twiddy et al., 2002), this implies that different serotypes are subject to different selection pressures even though the reasons are unclear. The above-mentioned researches were conducted in the most studied protein: the envelope protein.

Another point of view from DENV comparative genomics focusing on the detection of genomic recombination (which is a kind of genomic evolutionary feature), has been previously analyzed. The study of (Tolou et al., 2001) through rigorous sequence analysis provided strong evidence for the occurrence of intragenomic recombination events between eight DENV-1 compared genomes belonging to different lineages. Another comparative analysis of DENV also with focus on detecting events of recombination (Carvalho et al., 2010) comparing 42 genomes of DENV-1 and found three detectable intra-American DENV-1 recombinants. Recombination events might have major implications for virus evolution, pathogenicity, vaccine safety and efficiency, or diagnosis.

The analysis of codon usage is a scarce evolutionary field of study for DENV. Although these viruses have been previously studied in the context of genus *Flavivirus*

(Jenkins et al., 2001), RNA type viruses (Jenkins and Holmes, 2003), or DENV genomic comparisons (Behura and Severson, 2013; Ma et al., 2013; Ming-Wei et al., 2007; Zhou et al., 2013). These studies have provided important information; but; only a limited number of genomes were employed for their analyses.

Molecular evolutionary studies have long relied on the availability of genomic sequence data. The increase in the number of completely sequenced genomes has facilitated to address a number of fundamental questions about deep evolutionary relationships among the proteins encoded in virus genomes and by inference, among genomes themselves. Several efforts are currently underway to understand the global population structure of DENV and its genomic correlation with severe disease. The Broad institute (http://www.broadinstitute.org/annotation/viral/Dengue/GlobalPopulationStructure.html) initiate in 2005 a construction of a database with the aim to sequence >3500 dengue genomes of distinct geographic origin and disease to build the genomic infrastructure needed to study and combat the virus. This database contains at present 3199 complete genomes of DENV with information on the related disease and the isolation time. The information generated for The Broad institute is also deposited in GenBank (http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/).

The increasing number of genome sequences reported from Mexico and all over the world could thus help to reveal how DENV genomes diverge and what are the principal contributing factors for their evolution.

2.2.7. DENV molecular patterns associated with their pathogenesis

2.2.7.1. Genomic Signatures

Regarding to the more severe diseases previously commented, numerous epidemiologic studies had observed that severe dengue disease (DHF/DSS) is more often associated with secondary dengue infections (Kliks et al., 1988). The answer for development to the severe forms dengue disease could be deciphered thorough the analysis of genomic signatures of the virus proteins. In this respect some genomic comparative analyses on DENV has been achieved.

A research published in 1999 (Leitmeyer et al., 1999) compared full genome sequences of 11 dengue viruses of serotype 2 and found several structural differences between those viruses associated with DF only and those viruses with the potential to cause DHF: a total of six encoded amino acid charge differences were seen in the prM, E, NS4B, and NS5 genes while sequence differences observed within the 5' nontranslated region (NTR) and 3' NTR were predicted to change RNA secondary structures. The approach used in the study on dengue virus genome differences allowed their association with pathogenesis specially the aminoacid 390 forn the Envelope protein that was associated with the development of DHF (Leitmeyer et al., 1999) .

Another more recently comparative analysis realized manually on Envelope protein of several flaviviruses genera including DENV serotypes, revealed that, at the position corresponding to the glycosylated Asn-67 in dengue virus, the asparagine (Asn), is present in all compared viral species that cause hemorrhagic disease in humans (Barker et al., 2009). Thus for this pathogen, it is crucial to identify the molecular elements that are involved in the

disease development. Hence, the analysis of genomic signatures is important to understand the serotype determination and disease development.

2.2.7.2. DENV immunopathogenesis

The main target cells of DENV infection are monocytes, macrophages, dendritic cells and CD4 + and CD8 + lymphocytes. *In vitro* studies has been reported that DENV is capable to infect endothelial cells, liver cell lines, fibroblast and nervous cells, this could explain the syndromic behavior of dengue diseases (Modis et al., 2003).

During the primary infection, the antigenic determinants of the infected cells are presented to the lymphocytes via the MHC, it triggers an inflammatory process in the whole body (Libraty et al., 2001). In the course of the immune responses, the T cells release IFN alpha and gamma that stops the viral replication, and then, the NK cells are activated by cytokines that elicits a cytotoxic response that decline the infectious process. In addition, the release of large amounts of ILs and TNF prompts a pro-inflammatory defense on free viral zones and anti-inflammatory response in areas where the infection is active to limit tissue damage (Libraty et al., 2002). These factors regulate a correct immune response to achieve the viral clearance and only allow the presentation of DF clinical form.

Once the infection is controlled and has spent at least 7 days of antigen presentation to T cells, their counterparts (B lymphocytes) develop and begin to secrete IGs to form the humoral memory response and a second infection. In secondary infection by the same serotype or one with more than 70% homology, the immune response is immediately activated and release pro-inflammatory cytokines for viral control, which is efficient and

allows stopping the infection from the start, even without symptoms occur (Durán et al., 2010). However, if the infective virus is a new serotype (type 2 generally lack homology to the other), or which is not genetically homologous (within 70%, mainly due to mutations in the viral genome) the immune response is activated but this is not effective and it ends up being more damaging this latter because of two fundamental processes (Chaturvedi et al., 2000): The secondary responses activate the release of inflammatory cytokines (IL-4, IL-6, IL-8 and IL-10) by lymphocytes, leading to an increase of vascular permeability favoring the development of dengue hemorrhagic fever. Aforementioned viral immune complexes are captured by monocytes and macrophages which destroy the virus in phagosomes, however, the virus escapes (just as it happens in the endosome) generates destruction of immune cells and the spread of virus to other tissues due to increased migration presented by these cells. These two elements generate the pathophysiological basis for the development of dengue fever and while one can spread the virus, the other induces a state proinflammatory clotting activation (causing a CID for dengue shock) and bleeding (by blood extravasation) making this virus deadly in susceptible individual (Noisakran and Perng, 2008).

This has led to the hypothesis of antibody dependent enhancement (ADE) of DENV infection in which antibodies from a primary infection are able to recognize, but do not neutralize, the virus during a secondary infection, and the virus-antibody complex gains an entry into target cells via the Fc receptor (Burke and Kliks, 2006; Halstead, 2003).

3. JUSTIFICATION

Dengue virus is a serious public health problem in Mexico and in the world. The four serotypes of dengue viruses have been confirmed in the most parts of our country. However, only a small number of viral sequences from Mexico have been performed. Thus, providing new viral genome sequences and making genomic comparison at large scale will help us to understand globally the genetic features and evolutionary trends of dengue virus populations in Mexico and in the other regions of the world.

4. HYPOTHESIS

Dengue virus nucleotide and amino acid sequences contain the information that determines functional adaptation and their evolutionary history.

5. OBJECTIVES

5.1. General objective

To perform a large-scale sequence analysis using the tools of comparative genomics helps to understand the evolutionary mechanisms and adaptation to the host of dengue virus.

5.2. Specific objectives

1. To analyze the codon usage patterns of dengue virus genomes for the four serotypes
2. To identify genomic signatures that could be related to disease
3. To identify the epitope structures that could distinguish the four serotypes and could be related to distinct the immunology response
4. To analyze the genome sequences of dengue virus isolated in Mexico and perform comparison with the genomes published in public databases

6. MATERIALS AND METHODS

6.1. DENV culture in laboratory and viral fragment sequencing

6.1.1. Cell infection and RT-PCR

The C6/36 cell line was maintained in Eagle's Minimum Essential Medium (Sigma) supplemented with 10% fetal bovine serum (BioWest) and cultured at 28 degree with 5% carbon dioxide. When the cell line was confluent, the subcultures were performed by scraping or pipetting to new flasks. The yuc17438 DENV 2 virus stock, isolated from a patient 47 years old male from Yucatan México with DF, was provided by Dr. Isabel Salazar. 10 µL of the virus stock was used to infect the C6/36 cell lines. The infected cells were cultured for 7 days in the 28 degree incubator. The infected cells were harvested by pipetting and centrifuged at 4 degree (2000g for 10min). The supernatant and the pellet were used for dengue 2 virus RNA extraction respectively.

RNA was extracted from infected cells C6/36 using Qiagen for supernatant and Trizol reagent for the cell pellets according to the manufacturer instructions. The complementary DNA (cDNA) was produced using a reverse transcription system according to the PROMEGA company instructions, as follow. cDNA was synthesized using in a 0.2mL RNAse-free microtube, first 9.5 µL of viral RNA was incubated at 70 ºC for 10 min, then was briefly centrifuged and placed on ice. While a 10µL reaction was prepared, as follows: 1µL of 25mM genome specific synthetic oligonucleotide primer, 2µL of 10mM dNTP mix, 2µL of 10× RT buffer (PROMEGA), 4µL of 25Mm MgCl2, 0.5 µL RNasin ribonuclease inhibitor, and 1µL AMV RT enzyme at high concentration (15 U/µL) was added to the RNA mixture on ice for a final volume of 20µL. The resulting mix was incubated at 25 ºC for 10 min, then incubated at 42 ºC for 40 min. and heated at 95 ºC on Dyad Thermocycler (Peltier thermal

cycler). The cDNA product was then chilled on ice for 5min. The First strand cDNA products were stored at −80 ºC or used immediately in PCR assays.

6.1.2. PCR assays

PCR assays were performed in a 25 µL of reaction system containing 5 µL of cDNA in PCR reaction buffer, with 0.65 µL of 50 mM of MgCl2, 0.5 µL of 10mM dNTP mix, 2.5 µL of 10× PCR buffer, 1µL of 25mM of each primer, and 1.25 U of Taq DNA polymerase (Perkin-Elmer). PCR amplifications were implemented in a Dyad Thermocycler (Peltier thermal cycler) with the following conditions: an initial denaturation at 95 ºC for 2 min; 30 cycles of denaturation for 30 s at 95ºC, annealing for 30 s (55–60 ºC, according to the different Tm value of each pair of primers) and extension at 72 ºC (30 s – 4 min, depending on the length of the target products); and a final extension at 72 ºC for 10 min. The E amplification PCR products were separated in 1% agarose gel and visualized under UV light after SYBR Green ® stain (Figure 3)

6.1.3. Sequencing

The BigDye Terminator v3.1 Cycle-Sequencing Kit was used for cycle-sequencing reaction according to the manual. The program for the reaction is 96 ºC for 1 min; then 25 cycles at 96 ºC for 10 s, 50 ºC for 5 s, 62 ºC for 4 min; and 62 ºC for 1 min, then 4 ºC for maintenance.

The BigDye Terminator TM Purification Kit was applied for cleaning the cycle-sequencing reaction products. After the incubation at 25 ºC for 30 min, the mixture was centrifuged at 14,000 rpm for 2 min. The samples were used for sequencing in ABI 3130 automated sequencer. The chromatograms obtained from sequencing were assembled into

overlapping contigs using GeneStudio software version 2.2.0.0 (www.genestudio.com) and edited manually when the contigs were aligned to sequences of other known genomes of the same serotype.

6.2. Bioinformatic analysis

6.2.1. Gene and Genome sequences

The E gene accession numbers from Mexican sequences were obtained from (Carrillo-Valenzo et al., 2010; Diaz et al., 2006) and searched on the NCBI database along with the new reported Mexican sequences. The whole genome sequences of 3047 DENV 1-4 were downloaded from the NCBI DENV resource at: http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/. This website provided DENV information that includes sample sequence, location, serotype, etc. (Resch et al., 2009). Four datasets that correspond to each one of the four serotypes were established. They included 1336 genomes for DENV1, 927 genomes for DENV2, 670 genomes for DENV3, and 114 genomes for DENV4. The coding sequences of genomes were collected in a dataset for each serotype orderly according to their geographic regions of isolation as Africa, Asia, North America, Oceania, and South America and for the samples from the same continent along with the order of host sources as human, mosquito, monkey and unknown host. A number was then assigned to each genome in each dataset, which facilitates the subsequent analyses.

6.2.2. Nucleotide compositions and codon usage bias

The total GC% (GC) and GC% at 1st (GC1), 2nd (GC2) and 3th (GC3) codon position of coding sequences for each DENV genome sequence were calculated in order to show the

impact of selection on codon usage of DENV. The total GC content was calculated with the following equation:

$$GC = \frac{(G + C)}{(A + T + G + C)}$$

Where the G, C, A and T are the number of nucleotides in the genome. For the calculation of GC at the three codon positions we used the following equation:

$$GCn = \frac{Gn + Cn}{\left({}^{L}/_{3}\right)}$$

Where G*n*, C*n* are the number of guanines and cytosines at the n*th* (1, 2 or 3) position of the codon and *L* is the length of the genome.

Relative synonymous codon usage (RSCU) (Sharp and Li, 1987) was estimated as a proportion of the observed occurrence of codons to the expected occurrence when all codons for the same amino acid are equally used. The RSCU was calculated with the following equation.

$$RSCU = \frac{Xij}{\sum_{j}^{ni} Xij}\ ni$$

Where X_{ij} is the observed number of the *i*th codon for the *j*th amino acid which has n_i kinds of synonymous codons. It was measured for 59 codons except Met, Trp and the three stop codons for each genome tested in this study. Effective number of codons (ENC) is a parameter to reveal the number of equally used codon that could yield the observed codon

usage bias in a gene or a genome (Wright, 1990). ENC was calculated with the following equation.

$$ENC = 2 + \frac{9}{\overline{F}\,2} + \frac{1}{\overline{F}\,3} + \frac{5}{\overline{F}\,4} + \frac{3}{\overline{F}\,6}$$

Where $\overline{F}\,k$ (k = 2, 3, 4, 6) is the mean of $\overline{F}\,k$ values for the k-fold degenerate amino acids. ENC's values range from 20, the strongest bias, to 61, no bias. Because the genomes of DENV have unique uninterrupted polyprotein ORF, we applied ENC to quantify the level of codon usage bias on genome level in the present study. ENC prime (ENCp) was also used to quantify the codon bias taking into account the nucleotide background of the genomes (Novembre, 2002). The GC at three codon positions and RSCU were calculated with package seqinr (Charif and Lobry, 2007) for R (R-Development-Core-Team., 2010) and ENC and ENCp was calculated with the software Codonw and ENC prime, respectively (Novembre, 2002; Peden, 1999).

6.2.3. Correspondence Analysis

Correspondence Analysis (CA) is an effective method to show the relationship among multiple categorical variables by a statistical procedure. The unique condition is to have a non-negative data ordered in a two-way table for analysis. It is much better if the table consists of large enough dataset and homogenous variables (Greenacre, 2007). The RSCU and Aminoacid dataset form a table that should meet well the CA conditions. The RSCU and the Aminocid tables was read and formatted as data.frame in order to perform the CA with the

function "*dudi.coa*" using the ADE-4 package (Dray and Dufor, 2007) in R. In the results obtained, each genome was represented as 59-orthogonal axes (20-orthogonal axes for aminoacids), and each axis corresponds to one of 59 codons or 20 aminoacids. Thus the results of CA show how much DENV genomes are correlated to the level of codon usage or aminoacid variation patterns. The advantage of CA is that the results can be depicted as a map, in which each row and each column is represented as a point, which facilitates to understand the relation of codon usage and aminoacid bias among the genomes.

6.2.4. Evaluation of influencing evolutionary factors of DENV codon usage

Correlation analysis among GC1, GC2, GC3, GC, ENC, ENCp values and the selected axis of variation among each DENV1-4 dataset was performed, using the Pearson's rank correlation method. For better explanation of the correlation results only the coefficient ≥0.70 was considered as strong correlation (Suzuki et al., 2008; Taylor, 1990). As regards the evaluation of correlation coefficients, the null hypothesis of no correlation between the variables was tested at significance level of p=0.01.

6.2.4. Hierarchical Clustering based on codon usage

A distance matrix that accounts for differences in RSCUs for DENV genomes was constructed with the function "*dist*" and the Euclidean distance method by the software R. The matrix obtained were then used to aggregate the RSCU values of each genome sequence into hierarchical clusters of similar codon patterns with the function "*hclust*" and the ward method by the software R. The hclust objects produced, were then transformed to a phylo

objects for plotting the final trees with the ape (Paradis et al., 2004) and pyloch packages for R.

6.2.5. Gene, Genome and protein alignments

Alignments based on the nucleotide and protein sequences of genomes were done with the software MAFFT (Katoh and Standley, 2013). We used the default "*--auto*" function to run all the alignments on MAFFT.

6.2.6. Phylogenetic trees

The FastTree (Price et al., 2010) software was used to construct approximately-maximum-likelihood phylogenetic trees for each of DENV1-4 from whole alignments data. FastTree software can handle large alignments in a practical amount of time and memory. The generalized time-reversible (GTR) model was used for phylogenetic tree construction. To estimate the local support values of each split in the tree, the Shimodaira-Hasegawa test was used. The *Newick* tree files generated were used with the ape and pyloch packages for R to plot the phylogenetic trees. As the recombination has also impact on the evolution of DENV (Worobey and Holmes, 1999), we also tested this pattern using the software Recombination Analysis Tool (RAT) (Etherington et al., 2005) for each DENV dataset.

6.2.7. Aminoacid signature analysis

The whole DNA genome alignments for DENV1-4 datasets were separated by gene regions, and then each gene dataset was translated to aminoacids with the software seaview

(Gouy et al., 2010). The aminoacid proteins datasets obtained were searched by aminoacid polymorphism with the function “fasta2genlight” of adegenet package for R. Once the polymorphisms were extracted they were separated by continental regions and inspected with the weblogo tool (Crooks et al., 2004).

6.2.8. Epitope analysis

The epitopes of DENV were extracted from the Immune Epitope DataBase (IEDB) (Kim et al., 2012; Zhang et al., 2008). All epitopes retrieved from IEDB have been confirmed by experiments. The present study focuses on B and T-cell epitopes identified in situations of DF and DHF. All sequences of each viral protein were used to search for epitopes with the epitope conservancy analysis tool provided by IEDB. The shortest analyzed sequence of epitope was limited to be ≥ 5 residues for linear epitopes and ≥ 4 residues for discontinuous epitopes. After obtaining the required B and T-cell epitopes the conservation of epitopes on specific protein among the groups of strains were manually implemented. Only the epitopes with 100% identity were considered.

7. RESULTS

7.1. PCR and Sequencing

The use of the general primer d2a5B for reverse transcription designed for DENV-2 serotypes (Christenbury et al., 2010) enabled us to obtain the whole genome cDNA of the strain tested in this study (Figure 2). From the cDNA the PCR amplification with primers designed for Mexican DENV2 strains facilitated us to obtain three overlapping amplicons of the envelope gene region (Figure 2).

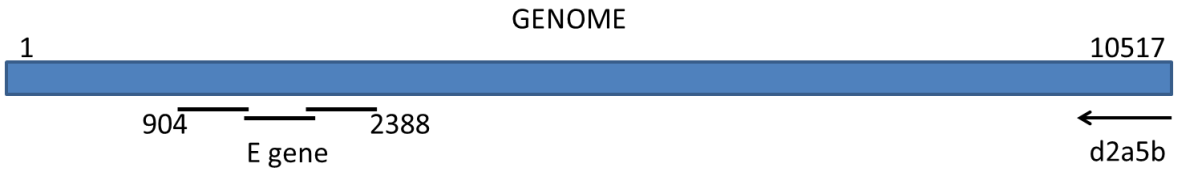


Figure 2. Schematic representation of the E gene location and the general reverse primer used for RT-PCR. The numbers at the extremes of the black line indicates the genome region of the E gene that is represented in the three overlapping amplicons, the black arrow represents the direction of the d2a5b primer used to amply the whole genome.

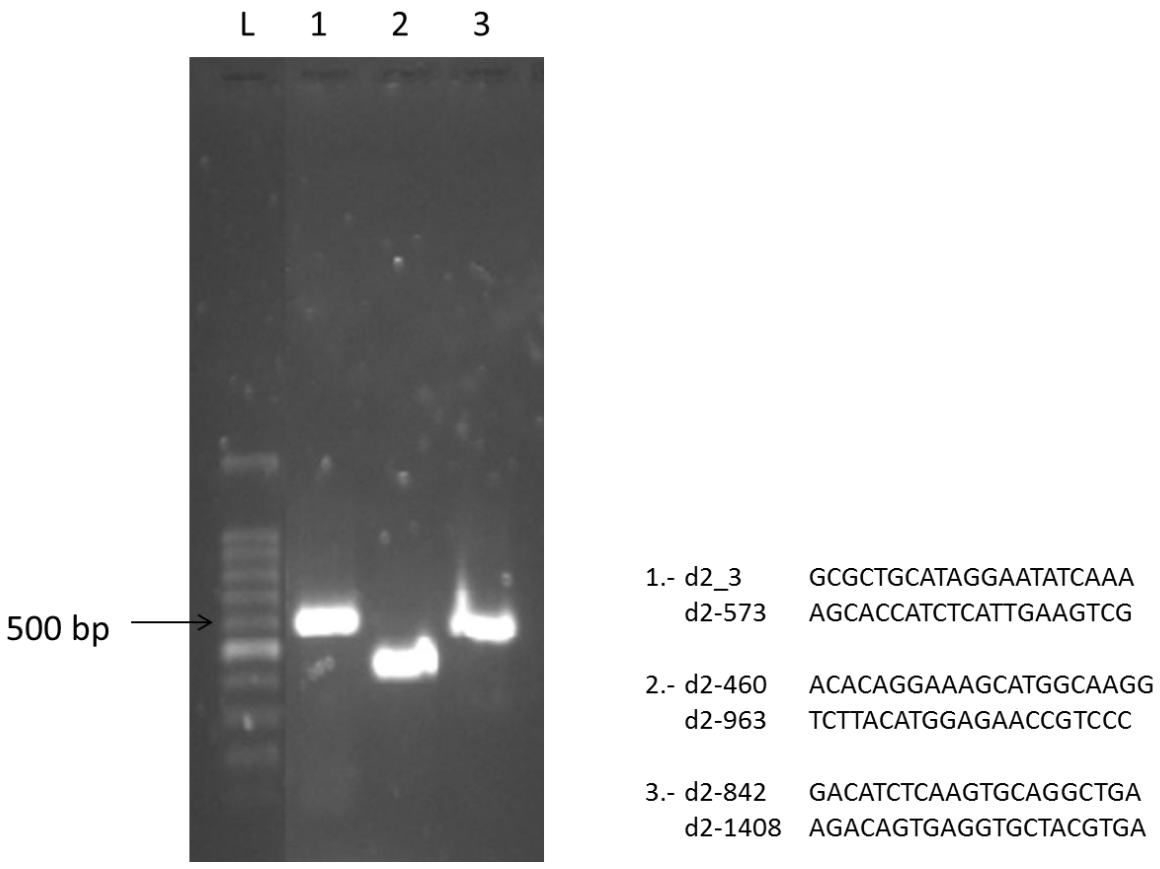


Figure 3. PCR amplification products for E gene DENV2 genome. L indicates the 100 bp DNA ladder, the numbers on the columns indicates the combinations of primers used and they are showed in the right side of the figure

The three overlapping PCR products (Figure 3) were sequenced, assembled and the curated nucleotide results were analyzed first by nucleotide blast analysis confirming that the DNA obtained was from a DENV2 serotype. Moreover, the blast result showed the close genetic relation of this Mexican strain with a New Guinea DENV 2 strain with 99% of identity (Appendix 1).

7.2. E gene phylogenetic analysis for new sequences reported from Mexico in the NCBI

The Bayesian phylogenetic analysis for DENV 1 revealed that the new sequences belongs to the genotype III, this result is very similar to a previous study. In the phylogeny, we observed the three introductions previously described and evidenced by the grouping pattern (Figure 4). The new reported sequences with isolation time of from 2008 to 2010, showed the absence of new DENV1 subtype introductions in the country.

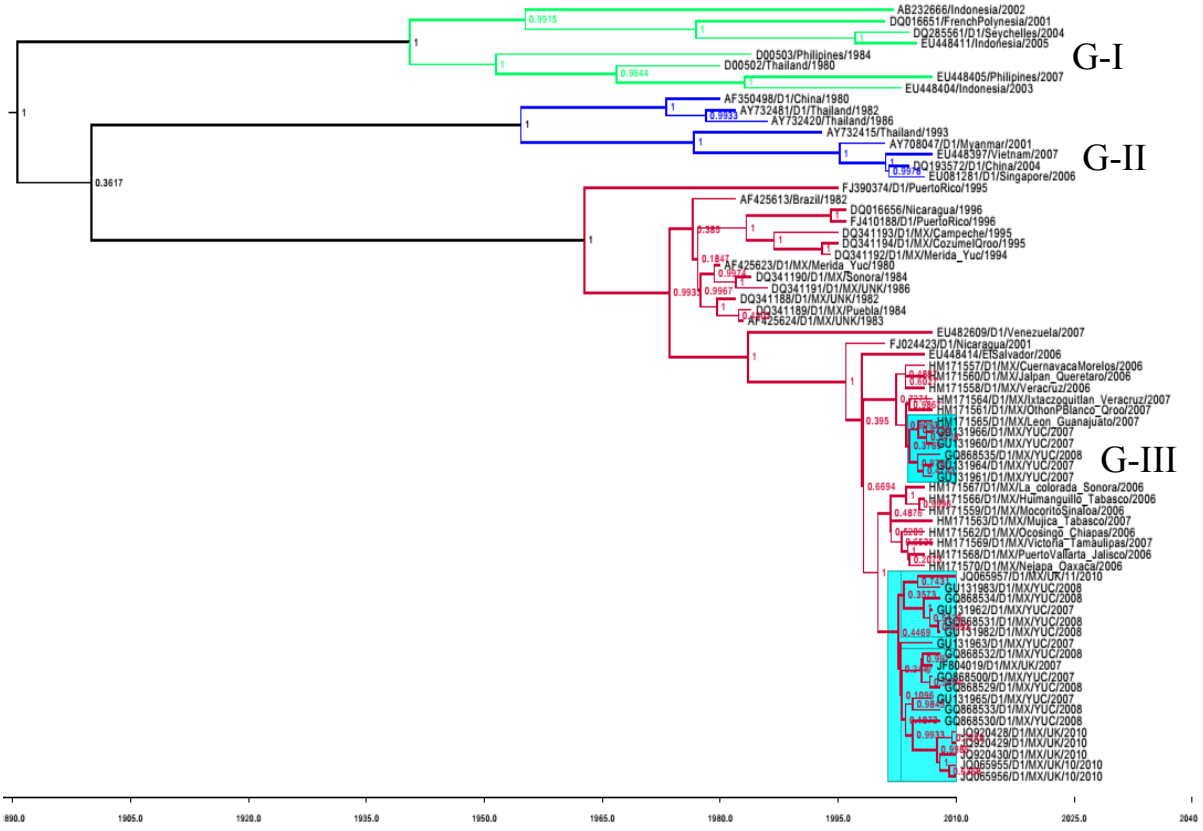


Figure 4. Phylogenetic analysis for DENV1 using representative E Mexican sequences. The cyan shaded branches indicate the sequences reported and not analyzed in the previous studies. The legends in front the tips indicates the genotype.

Using gene E of Mexican strains (DENV2) downloaded from NCBI database together with the sequenced E gene from DENV2 of this study (isolated in 2007 in Yucatan, Mexico), the bayesian phylogenetic analysis revealed that this new sequence DENV-2 belongs to the Asian 2 genotype (Figure 5), and closely related to a couple of Mexican strains from the Guerrero state. This genotype for DENV-2 could be introduced to Mexico in 1940s due to the close relation with the New Guinea strain isolated in 1944. In general, the phylogeny for DENV2 did not show new genotype introduction as in the last study and the new reported sequences belong to the Asiatic-American genotype.

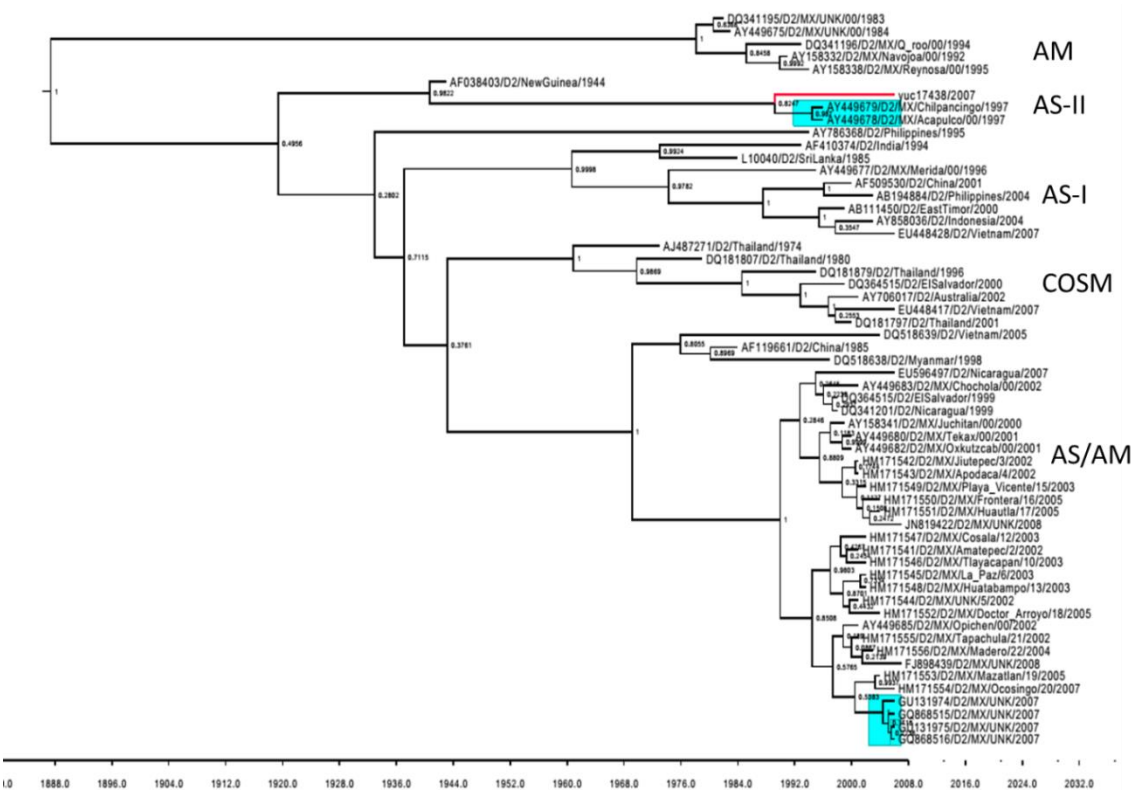


Figure 5. Phylogenetic analysis for DENV2. The red branch indicates the sequence obtained in this study. The cyan shaded branches indicate the sequences reported and not analyzed in the previous studies. The legends in front the tips indicates the genotype as follows AM (America), AS (Asia), Cosm (Cosmopolitan) and AS/AM (Asiatic/American)

For the analysis of the representative isolates from Mexico and two new DENV3 sequences isolated in 2007 showed that those sequences belong to the genotype 3 (Figure 6). This genotype is the current serotype circulating in Mexico according to the previous phylogenetic analysis. Thus, there are not new genotype introductions for DENV3.

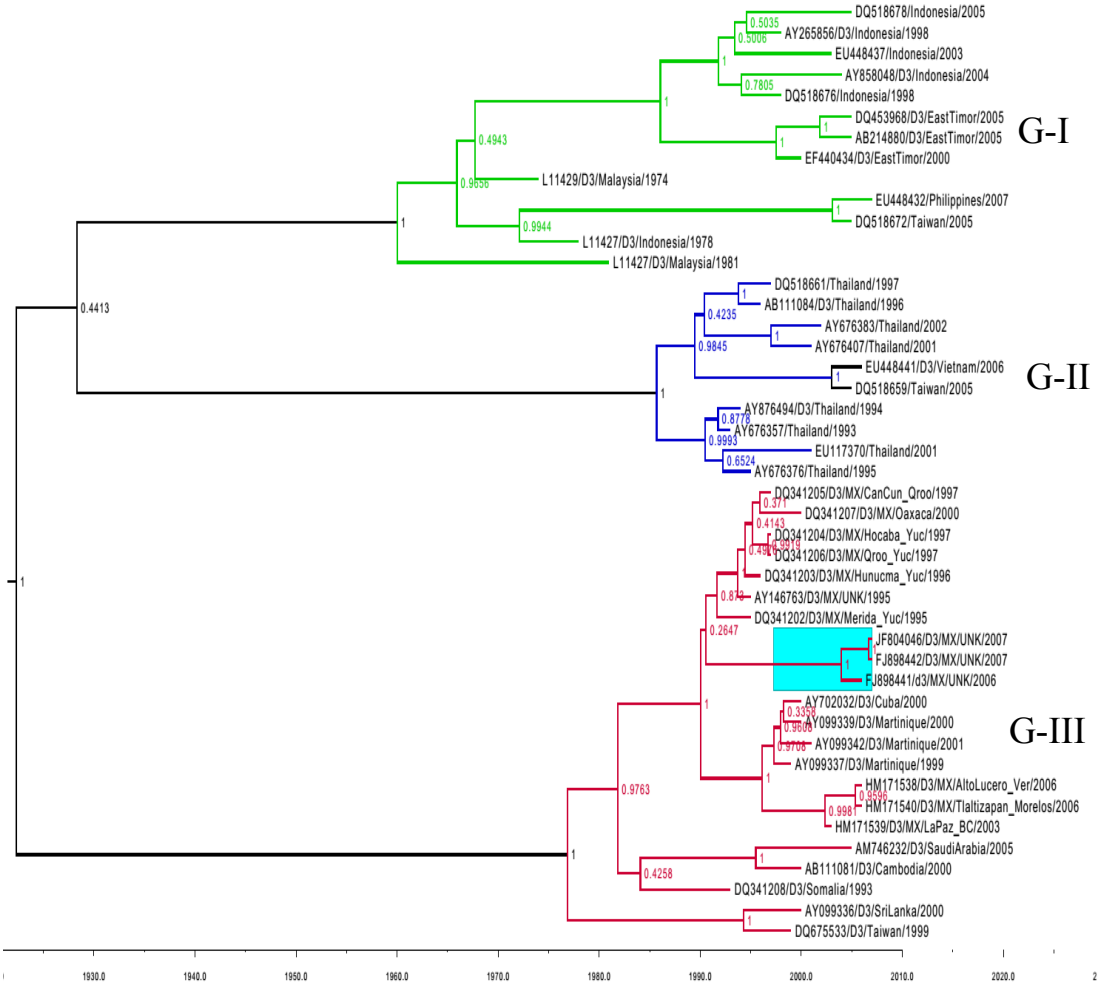


Figure 6. Phylogenetic analysis for DENV3. The cyan shaded branches indicate the sequences reported and not analyzed in the previous studies. The legends in front the tips indicates the genotype.

The phylogenetic analysis for DENV4 is shown in the Figure 7. We employ two 2006 isolates along with the previous reported sequences for the analysis. These two sequences were clustered within the genotype 2, which is the current circulating genotype for DENV4.

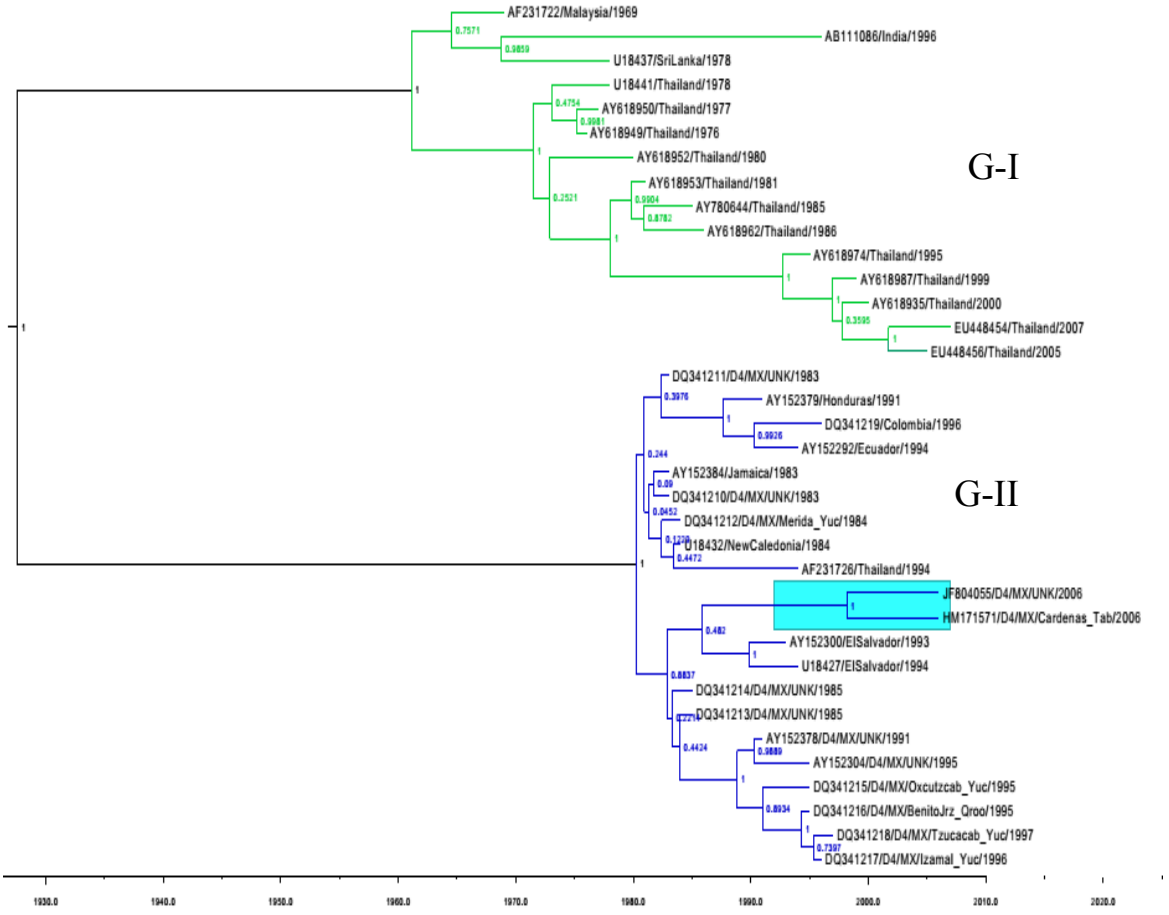


Figure 7. Phylogenetic analysis for DENV4. The blue shaded branches indicate the sequences reported and not analyzed in the previous studies. The legends in front the tips indicates the genotype.

7.3. Codon usage

The G＋C content and ENC. The overall G + C patterns at three nucleotide positions of codons were distinct for each DENV serotype (Figure 8A). The percentage of GC at the first nucleotide position of codon (GC1) is always the highest and GC2 is the lowest. GC1 in

Asia was the highest in comparison to the other regions. The total GC showed variability among the serotypes (Figure 8A) and a characteristic pattern was observed for each serotype. GC3 was more variable than GC1 or GC2 in general and its change was in expense of GC content at preceding positions, particularly GC2. The GC3 value was very close to the total GC and also showed high relationship to it (Appendix 3). DENV4 had higher GC3 than other serotypes. Meanwhile, the variation profile in GC content among genomes within a DENV serotype was apparently related to their geographic origin (Figure 8A).

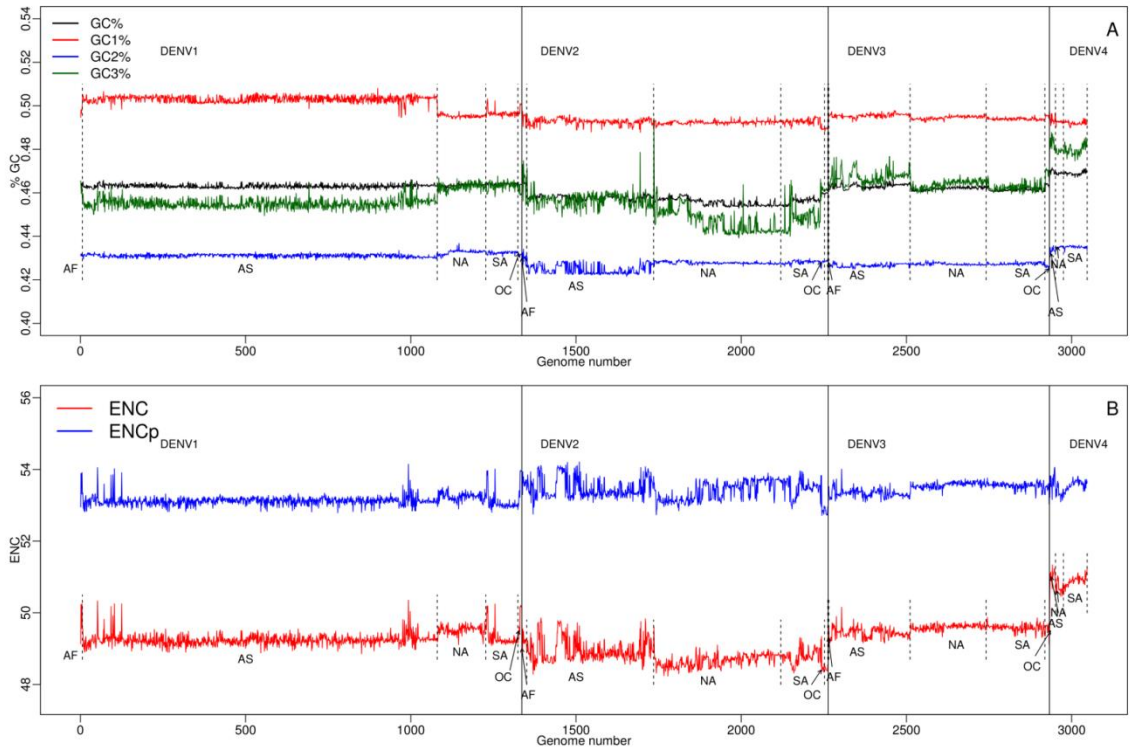


Figure 8. The nucleotide composition (G+C) and ENC, ENCp for the 3047 DENV1-4 genomes tested. A) Total GC and GC content at the three codon position for each genome. B) ENC and ENCp for each genome. The dashed lines in both figures indicate the geographical separation within a DENV serotype. The abbreviation means as follow: AF, Africa; AS, Asia; NA, North America; SA, South America; OC, Oceania.

A matrix correlation analysis with the total GC content and the GC at the three nucleotide positions of codons is shown in Appendix 3. The total GC content showed a strong correlation with the GC3 ($r \geq 0.7$, $p=0.01$) for DENV1-4. GC1 had also a strong correlation with GC2 and GC3 in DENV1.

The ENC was also analyzed for each serotype. DENV2 appeared with the highest codon bias with a mean 48.8 ± 0.28 whereas DENV4 showed the lowest bias mean (50.87 ± 0.17). ENC bias among genomes within a DENV serotype was correlated with their geographic origin (Figure 8B, line red). The ENCp analysis showed that the four serotypes had a homogenous codon bias in contrast to ENC (Figure 8B, line blue). A curve of ENC and ENCp values for each DENV 1-4 genome *v.s.* their corresponding GC3s data is shown in Figure 9 A, B. All points of the genome coding sequences lay below the predictable curve. The correlation analysis of ENC and ENCp showed almost no correlation with GC at any of the three codon positions for all DENV 1-4 (Appendix 3). These results indicate that, independent of compositional constraint, some other factors that affect the codon usage variations exist.

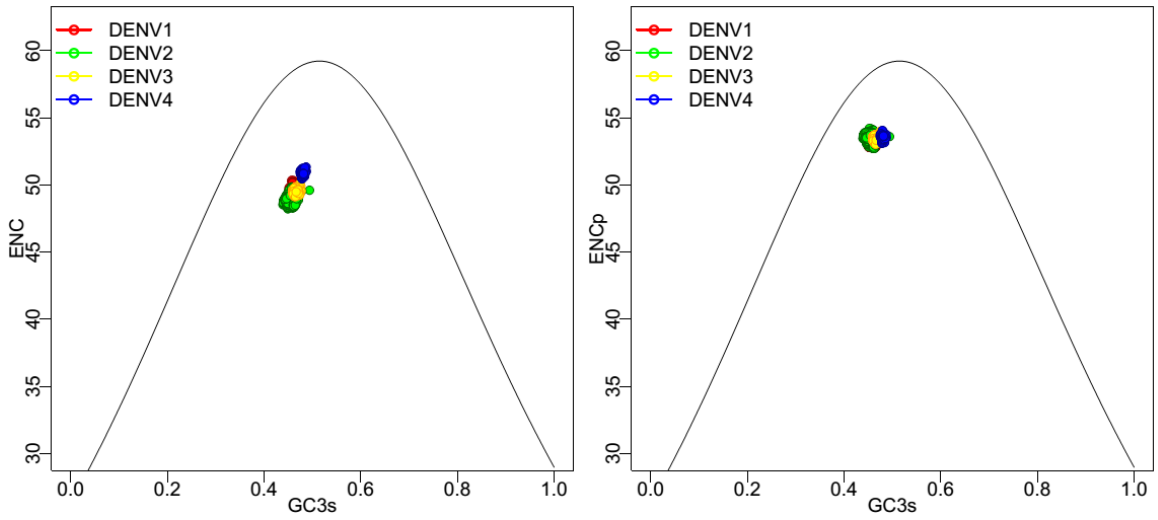


Figure 9. **Effective Number of Codons vs GC3s plot of genomes of DENV1-4.** A) ENC *vs* GC3s, B) ENCp *vs* GC3s.

Preferred codons. The mean and standard deviation of RSCU for 18 amino acids except Met, Trp, and stop codons were determined for each serotype (Table 1). Eleven preferred codons, AGA (Arg), AAC(Asn), GAC(Asp), GAA(Glu), GGA(Gly), ATA(Ile), AAA(Lys), CCA(Pro), TCA(Ser), ACA(Thr) and GTG(Val) were consistently shared for all the four DENV (highlighted in blue). There was no extreme bias in preferred codons among specific serotypes. Although in some cases we observed the preferred codons for specific serotypes, these codons still belonged to the set of codons mainly used for the other serotypes. For example, DENV1 used more commonly TAT (Tyr) codon instead of the preferred TAC (Tyr) by DENV2-4. The CAC (His), not CAT (His) codon was preferred in DENV1 and 3 while DENV2 and 4 use these two codons Tyr or His at proximate frequency. Codon CTG (Leu) was preferred by DENV1 and 2, but DENV3 and 4 preferred the codon TTG (Leu).

Table 1. RSCU for DENV 1-4

AMINOACID	CODON	DENV1	DENV2	DENV3	DENV4
Ala	GCA	1.41 ±0.05	1.56 ± 0.05	1.28 ± 0.03	1.16 ± 0.03
	GCC	1.28 ±0.03	1.15 ± 0.06	1.12 ± 0.05	1.29 ± 0.05
	GCG	0.36 ±0.03	0.31 ± 0.04	0.47 ± 0.04	0.39 ± 0.02
	GCT	0.94 ±0.04	0.99 ± 0.05	1.13 ± 0.04	1.17 ± 0.04
Arg	AGA	3.15 ±0.04	3.40 ± 0.08	3.31 ± 0.04	3.13 ± 0.08
	AGG	1.38 ±0.04	1.19 ± 0.07	1.56 ± 0.05	1.71 ± 0.06
	CGA	0.50 ±0.04	0.44 ± 0.04	0.31 ± 0.02	0.44 ± 0.04
	CGC	0.40 ±0.04	0.36 ± 0.06	0.33 ± 0.04	0.32 ± 0.03
	CGG	0.27 ±0.03	0.17 ± 0.06	0.22 ± 0.05	0.22 ± 0.04
	CGT	0.30 ±0.05	0.43 ± 0.06	0.27 ± 0.04	0.18 ± 0.03
Asn	AAC	1.12 ±0.06	1.01 ± 0.06	1.17 ± 0.04	1.18 ± 0.04
	AAT	0.88 ±0.06	0.99 ± 0.06	0.83 ± 0.04	0.82 ± 0.04
Asp	GAC	1.09 ±0.04	1.16 ± 0.04	1.12 ± 0.03	1.11 ± 0.03
	GAT	0.91 ±0.04	0.84 ± 0.04	0.88 ± 0.03	0.89 ± 0.03
Cys	TGC	0.95 ±0.05	0.97 ± 0.11	0.98 ± 0.05	1.07 ± 0.04
	TGT	1.05 ±0.05	1.03 ± 0.11	1.02 ± 0.05	0.93 ± 0.04
Gln	CAA	1.19 ±0.04	1.24 ± 0.06	1.29 ± 0.03	1.00 ± 0.04
	CAG	0.81 ±0.04	0.76 ± 0.06	0.71 ± 0.03	1.00 ± 0.04
Glu	GAA	1.21 ±0.03	1.4 ± 0.04	1.18 ± 0.01	1.25 ± 0.02
	GAG	0.79 ±0.03	0.6 ± 0.04	0.82 ± 0.01	0.75 ± 0.02
Gly	GGA	2.35 ±0.04	2.25 ± 0.06	2.11 ± 0.03	2.04 ± 0.04
	GGC	0.51	0.55 ±	0.63 ±	0.54 ±

		±0.02	0.04	0.03	0.04
	GGG	0.65 ±0.04	0.73 ± 0.03	0.85 ± 0.02	0.92 ± 0.04
	GGT	0.48 ±0.02	0.47 ± 0.03	0.41 ± 0.02	0.50 ± 0.04
His	CAC	1.13 ±0.04	0.98 ± 0.06	1.08 ± 0.09	1.00 ± 0.06
	CAT	0.87 ±0.04	1.02 ± 0.06	0.92 ± 0.09	1.00 ± 0.06
Ile	ATA	1.37 ±0.02	1.2 ± 0.05	1.31 ± 0.03	1.16 ± 0.03
	ATC	0.82 ±0.03	1.02 ± 0.05	0.74 ± 0.03	0.89 ± 0.05
	ATT	0.82 ±0.04	0.78 ± 0.04	0.95 ± 0.04	0.95 ± 0.04
Leu	CTA	1.36 ±0.13	1.06 ± 0.08	0.91 ± 0.09	0.86 ± 0.07
	CTC	0.63 ±0.04	0.93 ± 0.04	0.91 ± 0.05	1.00 ± 0.05
	CTG	1.51 ±0.07	1.52 ± 0.06	1.21 ± 0.07	1.29 ± 0.04
	CTT	0.70 ±0.03	0.64 ± 0.04	0.84 ± 0.04	0.71 ± 0.04
	TTA	0.68 ±0.09	0.7 ± 0.08	0.82 ± 0.04	0.76 ± 0.05
	TTG	1.11 ±0.10	1.14 ± 0.07	1.31 ± 0.05	1.38 ± 0.05
Lys	AAA	1.34 ±0.02	1.3 ± 0.04	1.16 ± 0.02	1.25 ± 0.01
	AAG	0.66 ±0.02	0.7 ± 0.04	0.84 ± 0.02	0.75 ± 0.01
Phe	TTC	1.06 ±0.04	1.09 ± 0.08	0.93 ± 0.06	0.78 ± 0.05
	TTT	0.94 ±0.04	0.91 ± 0.08	1.07 ± 0.06	1.22 ± 0.05
Pro	CCA	2.30 ±0.05	2.32 ± 0.04	2.18 ± 0.07	1.84 ± 0.03
	CCC	0.70 ±0.06	0.65 ± 0.09	0.74 ± 0.05	1.09 ± 0.06
	CCG	0.34 ±0.04	0.27 ± 0.05	0.18 ± 0.04	0.37 ± 0.02
	CCT	0.66 ±0.05	0.76 ± 0.07	0.9 ± 0.05	0.70 ± 0.05
Ser	AGC	0.83 ±0.05	0.92 ± 0.09	0.9 ± 0.07	0.71 ± 0.06
	AGT	0.73 ±0.04	0.99 ± 0.07	0.72 ± 0.06	0.81 ± 0.06

	TCA	2.22 ±0.05	2.02 ± 0.07	2.17 ± 0.06	2.08 ± 0.06
	TCC	1.11 ±0.06	0.88 ± 0.07	0.97 ± 0.05	0.89 ± 0.08
	TCG	0.26 ±0.03	0.37 ± 0.05	0.44 ± 0.07	0.42 ± 0.04
	TCT	0.85 ±0.06	0.82 ± 0.06	0.81 ± 0.05	1.10 ± 0.10
Thr	ACA	1.81 ±0.04	2.03 ± 0.03	2.13 ± 0.03	1.82 ± 0.04
	ACC	0.97 ±0.07	0.85 ± 0.06	0.74 ± 0.04	1.03 ± 0.03
	ACG	0.51 ±0.04	0.51 ± 0.03	0.48 ± 0.03	0.48 ± 0.03
	ACT	0.70 ±0.05	0.61 ± 0.06	0.64 ± 0.04	0.67 ± 0.04
Tyr	TAC	0.96 ±0.04	1.25 ± 0.07	1.14 ± 0.05	1.03 ± 0.03
	TAT	1.04 ±0.04	0.75 ± 0.07	0.86 ± 0.05	0.97 ± 0.03
Val	GTA	0.61 ±0.06	0.58 ± 0.06	0.59 ± 0.04	0.62 ± 0.04
	GTC	0.73 ±0.05	0.95 ± 0.08	0.91 ± 0.05	0.95 ± 0.05
	GTG	1.74 ±0.07	1.63 ± 0.07	1.64 ± 0.04	1.71 ± 0.03
	GTT	0.91 ±0.05	0.84 ± 0.06	0.85 ± 0.05	0.72 ± 0.04

*In blue color the shared preferred codons, in yellow color the preferred codons among specific DENV serotypes.

Correspondence analysis. One factorial axis accounted for 41.8%, 39.6% and 40.9% respectively of the total variability, indicating that one factor was predominant to explain the variability in DENV1-3 while for DENV4 dataset the first axis accounted for 25%. The first two axes accounted for more than half of that variability (53-56%) for DENV 1-3 except for DENV4 (41%). Thus, the first two factorial axes contribute to the principal differences in codon usage for DENV datasets.

The factor maps produced by crossing axes with the major sources of variation showed well-demarcated geographic separation. They exhibit the following features: 1) In the first axis, as the most important factor on the maps for each serotype (Figure 10 A-D), the genomes were divided into clusters accordingly to their geographic origin; 2) the Asian, African and Oceanic genome sequences tend to cluster together; 3) the North American and South American clustered together; 4) the Asian appeared more dispersed than those from other regions； 5) The genomes from other hosts (mosquito, monkey and unknown host) also clustered accordingly with their geographic sites of isolation. DENV2 showed the most complex geographic pattern. On the other hand, there were also some noticeable "outliers" (including three from Mexico for DENV2) in the figures, i.e. the genomes that were not located in the cluster with the majority of strains sharing the same geographic origin (Figure 10, Appendix 4). Some of these strains have been previously mentioned outside the general cluster in their phylogenetic analysis. For example, the genome from Djibouti Africa of serotype 1, previously observed more closely related to the Asian strains due to the existence of a recombinant sequence region with a strain from Singapore (Asia) (Tolou et al., 2001), was located on the Asian cluster in our analysis. Based on this finding we tested if the other identified outliers are recombinant strains with the software RAT. However, no sign of recombination was detected.

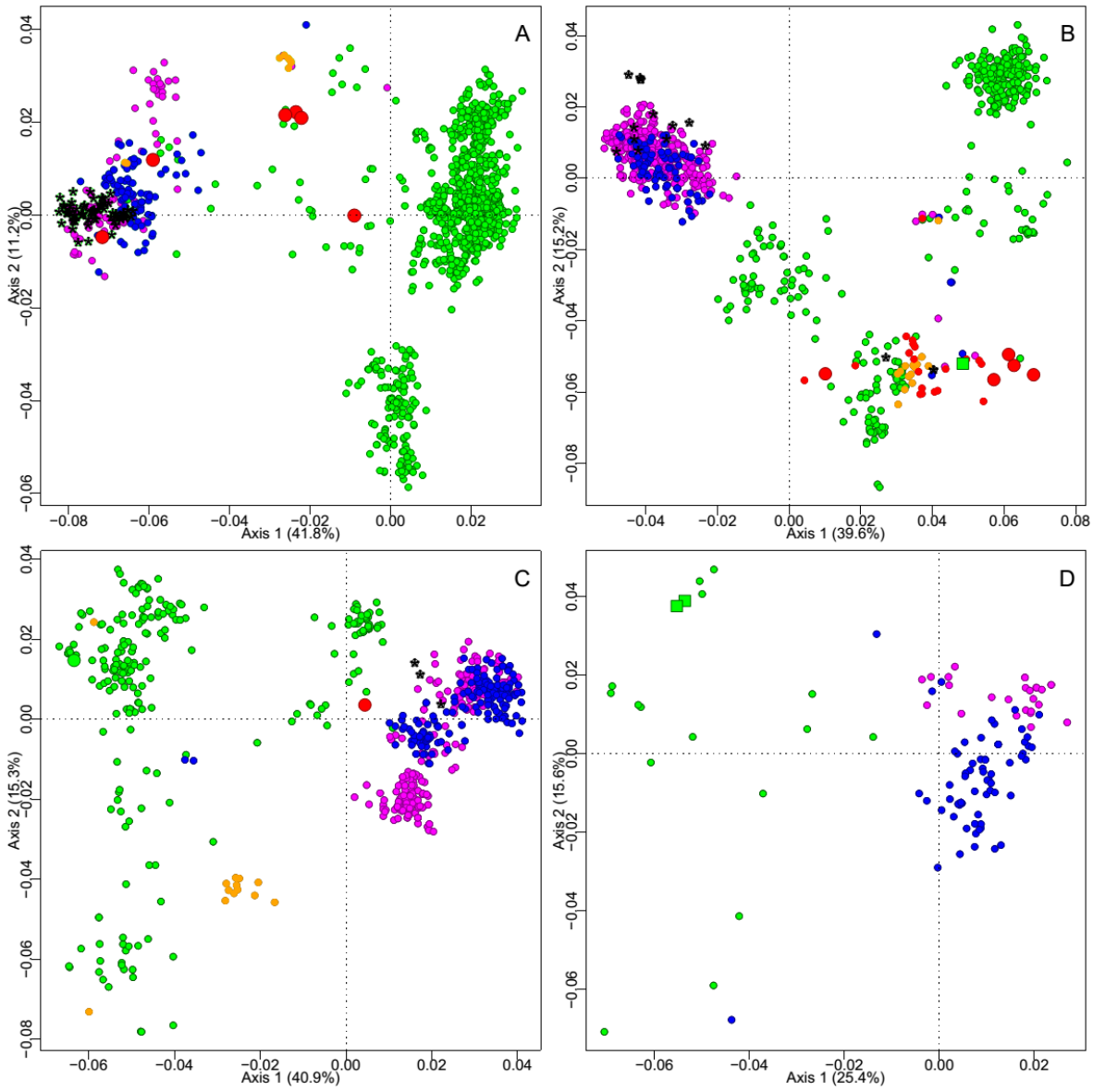


Figure 10. Correspondence analysis based on RSCU values for DENV. The geographic regions of isolates are indicated in colors as red (African), green (Asian), magenta (North American), blue (South American), orange (Oceanic) and black (Mexico). The host sources are respectively represented as circle for human, squares for monkey, inverted triangles for mosquito and asterisks for unknown host. A) DENV1; B) DENV2; C) DENV3; D) DENV4.

The correlations of the GC, GC1, GC2, GC3, ENC and ENCp of each genome with its position on the first axis are shown in Table 2. Depending on the specific serotype, the genome position on first axis had strong correlation with GC1 and GC3 for DENV1, GC2 and GC3 for DENV2 while others showed less correlation. It is interesting that GC3 showed negative relation with the first axis of the major variation for DENV1 but showed positive correlation with DENV2. ENC and ENCp showed no important correlation with all DENV 1-4.

Table 2. **The correlation analysis of GC, ENC and ENCp with the first axis of major variation**

Serotype	A1* % of variation	GC (r)	GC1(r)	GC2(r)	GC3(r)	ENC(r)	ENCp(r)
DENV1	41.8	-0.40	0.87	-0.54	-0.88	-0.47	-0.18
DENV2	39.8	0.57	0.00	-0.79	0.78	0.19	-0.18
DENV3	40.9	-0.55	-0.55	0.61	-0.59	0.44	0.59
DENV4	25.4	-0.40	-0.37	0.66	-0.46	-0.55	-0.50

* It represents the axis 1 in the correspondence analysis.

Phylogenetic and Hierarchical Clustering-based Trees for whole DENV genomes. The Hierarchical Clustering based Trees (HCbT) resulted from the RSCU data is shown in Figure 11 A-D. The HCbT revealed two major clusters in each serotype virus. The cluster consisting of Asian, African and Oceanic genome sequences tends to group together, whereas the cluster enclosing South and North American sequences assembles together. However, some Asian genomes were located at the cluster of North and South American strains. The genomes identified as outliers were also located in the same geographical clusters as indicated in our

CA analysis. On the other hand, the phylogenetic relationships among DENV genomes were also constructed based on the genome nucleotide sequences (Figure 11 E-H). The comparison of these phylogenetic trees showed that these analyses on two datasets produced similar results. Moreover, the majority of the outliers were also confirmed by the inferred phylogenetic trees.

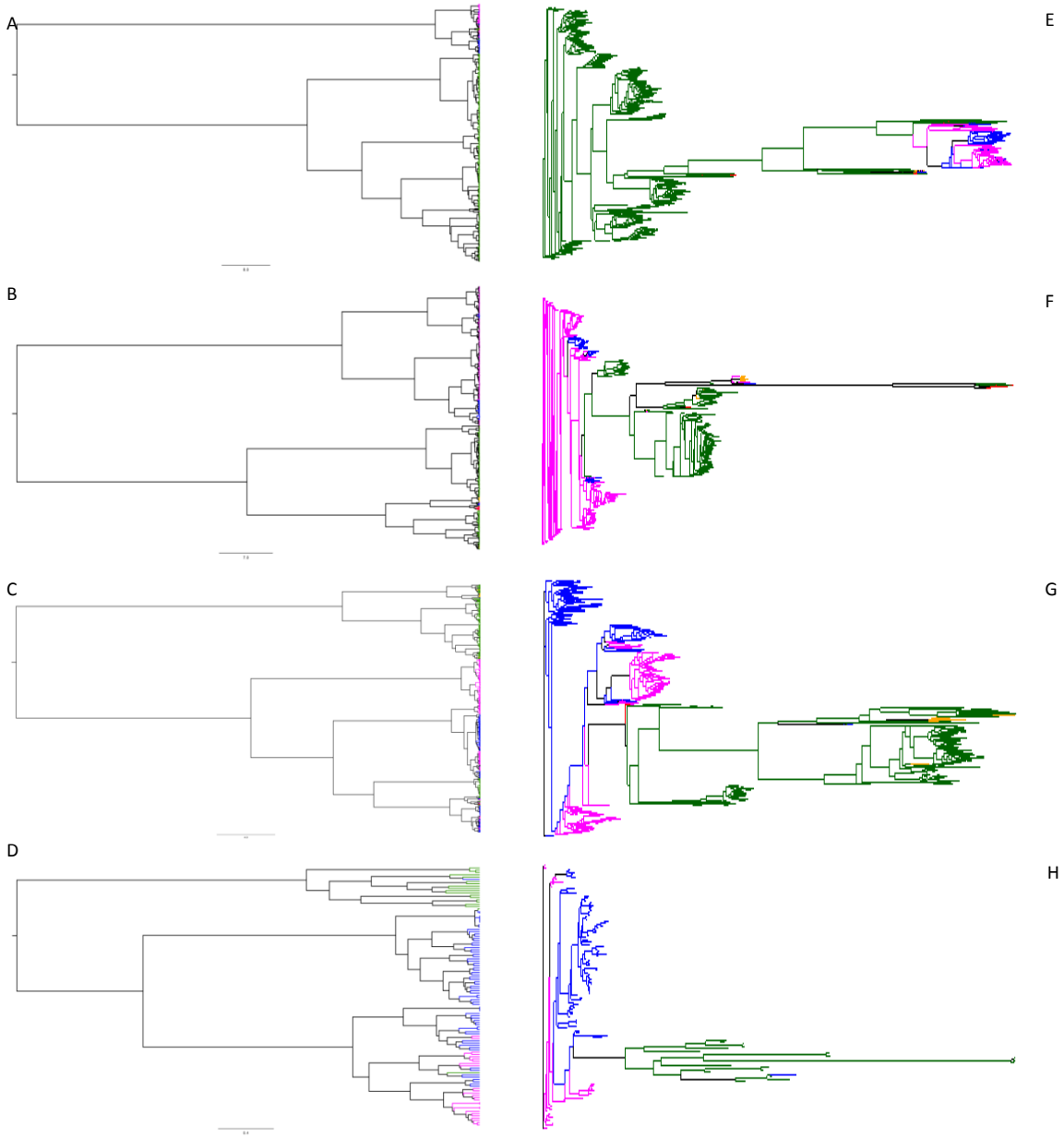


Figure 11. Hierarchical clustering trees based on RSCU data and Phylogenetic trees based on the genome nucleotide sequences. All the clades in each analysis of DENV1-4 were marked in different colors according to the source origin for better visualizing the phylogenetic clade of each genome representing as red (African), green (Asian), Magenta (North American), Blue (South American), Yellow (Oceanic), black (Mosquito) and Cyan (Monkey) . A-D) Hierarchical cluster based trees based on RSCU values for DENV 1-4, respectively. E-H) Phylogenetic trees based on the GTR nucleotide substitution model for DENV 1-4, respectively.

7.4. Amino acid signatures.

The analysis of the number of amino acid changes by protein is shown in Table 3. The serotype DENV 2 are more polymorphic by protein sequences than the other ones. It is interesting that DENV4 showed a similar amount of amino acid polymorphism as DENV1 and showed less amino acid changes than DENV2. DENV4 and DENV2 serotype showed the highest amino acid changes in the NS5 protein (72, 105). The proteins E (54), NS3 (53), NS1(48) and NS2A (51) also showed higher polymorphism in DENV2. DENV4 has less genomes compared but showed more variations than DENV1 and DENV3, particularly for those proteins E, NS1, NS2A, NS4A, NS4B, NS5, which diversified more in DENV4 than in DENV1 and DENV3.

Table 3. Aminoacid polymorphism by protein for DENV1-4

	C	M	E	NS1	NS2A	NS2B	NS3	NS4A	NS4B	NS5
DENV1	13	18	39	26	40	14	27	8	11	56
DENV2	14	28	54	48	51	14	52	17	24	105
DENV3	11	11	32	25	24	4	24	14	6	56
DENV4	10	14	53	44	43	10	24	15	18	72

Focusing in the E protein due to the known importance, we analyzed the amino acid sequence in more detail. The analysis of the polymorphic aminoacid with the help of the weblogo showed that for each serotype exists a clearly an amianoacid pattern related to the region of isolation (Figure 12) i.e. those changes correspond to a geographic signature. Furthermore, we observed that the aminoacid 390 from the E protein associated to the development of DHF (N D) (Leitmeyer et al., 1999) is poorly related to a geographical distinction (Figure 13 B).

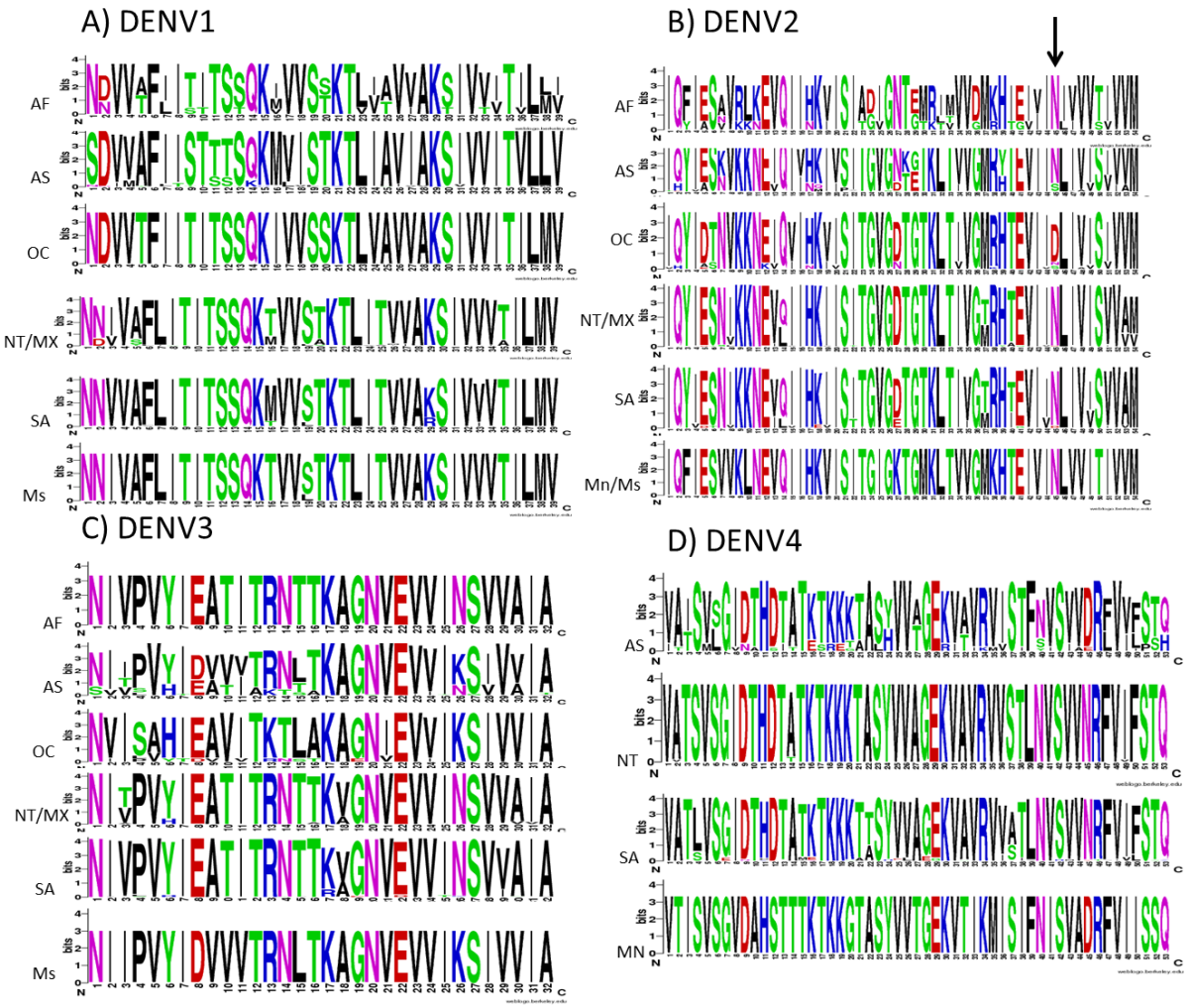


Figure 12. The logo presentation of aminoacid polymorphism for E protein from different regions within serotype. A-D) DENV1-4. The numbers under the logos represent the corresponding amino acid polymorphism. Each logo consists of stacks of letters, one stack for each position in the sequence. The overall height of each stack indicates the sequence conservation measured in bits, whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino acid at that position. The left letters indicates the source origin of the sequences. The arrow in the figure B indicates de aminoacid 390.

Due to the previous results we also analyze the aminoacid composition using the CA. The first factorial axis accounted for 53.8%, 41.1% and 51.9% respectively of the total variability, in DENV1-3 while for DENV4 dataset the first axis accounted for 38.1 %. The first two axes accounted for more than half of that variability (53-56%) for DENV 1-4. Thus, the first two factorial axes contribute to the principal differences in aminoacid usage for DENV datasets.

The factor maps produced by crossing the axes with the major sources of variation showed a well-demarcated geographic separation. Thus, they exhibit a similar feature as in the CA-RSCU analysis (though less dispersed than codon usage) : 1) In the first axis, as the most important factor on the maps for each serotype (Figure 13 A-D), the genomes were divided into clusters accordingly to their geographic origin; 2) the Asian, African and Oceanic genome sequences tend to cluster together; 3) the North American and South American clustered together; 4)The Asian appeared more dispersed than those from other regions； 5) The genomes from other hosts (mosquito, monkey and unknown host) also clustered accordantly with their geographic sites of isolation. DENV2 showed the most complex geographic pattern. Furthermore, the "outliers" were related as in the RSCU analysis.

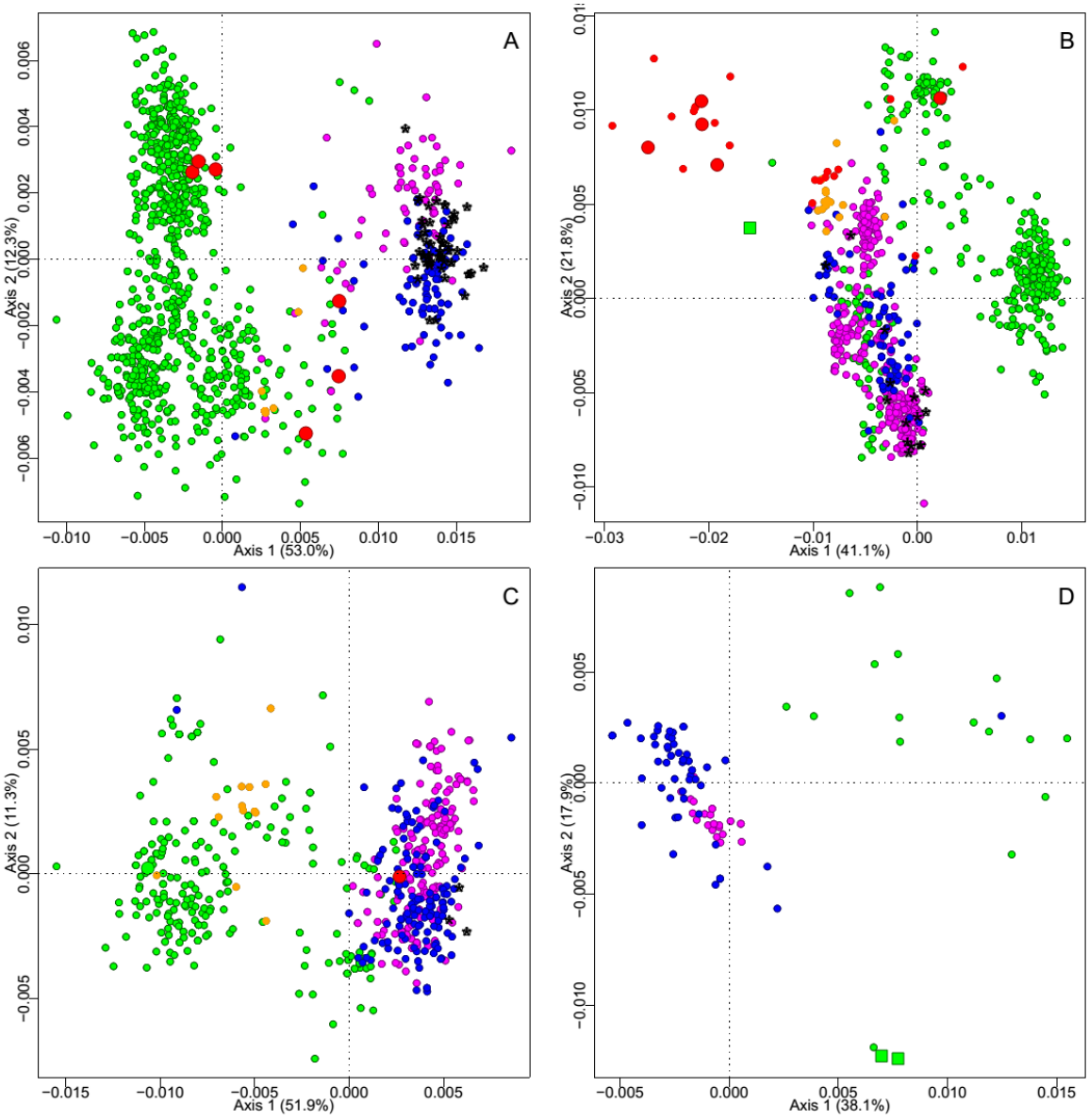


Figure 13. Correspondence analysis based on aminoacid frequencies for DENV. The geographic regions of isolates are indicated in colors as red (African), green (Asian), magenta (North American), blue (South American), orange (Oceanic) and black (Mexico). The host sources are respectively represented as circle for human, squares for monkey, inverted triangles for mosquito and asterisks for unknown host. A) DENV1; B) DENV2; C) DENV3; D) DENV4.

7.5. Epitope analysis.

The epitopes were collected from IEDB. Those epitopes were confirmed by experiments in IEDB. These epitope information in relation to dengue immune receptors by serotypes predominates for DF over DHF (Table 4) and further, B cell epitopes information is more related to DF while T cell epitope information is more related to the DHF. This observation has been previously reported in an epitope meta-analysis for flavivirus which include DENV (Vaughan et al., 2010).

Table 4. B cell and T cell epitopes by disease state available in IEDB

	B cell		**T cell**	
	DF	DHF	DF	DHF
DENV1	21*	1*	11	8
DENV2	252/16*	21/1*	27	52
DENV3	2*		14	4
DENV4	5/3*		3	4
Total	257/42*	21/2*	55	68

* Discontinuous epitopes

The epitope information is abundant on the E protein for both B and T cell. NS1 epitope information is predominantly for B cell and NS3 epitope information for T-cell. Those epitopes available for the whole DENV proteins were mapped on their corresponding protein sequence. The B and T cell epitopes that were conserved at 100% of identity of sequences are shown in table 5 and 6. The majority of the epitopes retrieved are well

conserved in the proteins (except for discontinuous epitopes). These conserved epitopes are more likely to be associated with the development of the disease (or protection) in all over the world isolates.

Table 5. B cell epitopes conserved with 100% of identity on DENV proteins associated with disease states

	C	prM/M	E	NS1	NS2A	NS2B	NS3	NS4A	NS4B	NS5
DENV1-DF			9*							
DENV1-DHF										
DENV2-DF	2		243	2				4		
DENV2-DHF				21						
DENV3-DF										
DENV3-DHF										
DENV4-DF			2*	4			1			
DENV4-DHF										
Total	2		243/11	27			1	4		

*Discontinuous epitope

Table 6. T-cell epitopes with 100% of identity on DENV proteins associated with disease states

	C	prM/M	E	NS1	NS2A	NS2B	NS3	NS4A	NS4B	NS5
DENV1-DF	1		1				9			
DENV1-DHF			1				6			
DENV2-DF		2	12	3	1	2	4	1		2
DENV2-DHF	3	3	8	1	1		34	2		1
DENV3-DF			3	2	1		5			3
DENV3-DHF					1		1			1
DENV4-DF			1				2			
DENV4-DHF							4			
Total	4	5	26	6	4	2	65	3	0	7

8. DISCUSSION.

8.1. Mexican E gene phylogenetic analysis

A study performed by Carrillo-Valenzo et al. (Carrillo-Valenzo et al., 2010) analyzed E gene sequences in order to trace the evolution of DENV Mexican strains collected until 2007. Since this research, more gene and genome sequences of Mexican strains have been reported in the NCBI database. To trace the molecular evolution of those sequences we took all the previous sequences for the present analysis.

The molecular analysis based on the nucleotide sequences from the recent strains compared with the previous strains can help us to understand the patterns and dynamics of the virus transmission in Mexico. A previous study found DENV-2—C932/Guerrero-Mx/97 and C1077/Guerrero-Mx/97 closely related to the New Guinea C (NGC) strain isolated in 1944,

this group belongs to the Asian 2 genotype (Diaz et al., 2006), we also identified the strain obtained in this study much related to NGC. In addition, the present analysis, using the most recent isolates from Mexico and based on the nucleotide of E proteins for DENV1-4, showed the similar results to (Carrillo-Valenzo et al., 2010) (Figures 4-7), indicating that there are not new introductions of new serotypes in the country. Thus we can infer that Mexican strains remain genetically stable confirming the previous comment that limited nucleotide variety exists in American DENV strains (Behura and Severson, 2013; Rivera-Osorio et al., 2011). These results confirm the subsequent analysis based on whole genomes (including the Mexican strains) described in a DENV global context.

8.2. Codon usage

The identification of principal factors shaping codon usage is important to understand the evolution of organisms, including viruses. The DENV1-4 global analysis of total GC relation with the three nucleotide positions of codon (GC1, GC2 and GC3) for whole genomes showed that the forces shaping codon usage were not the same for all codon positions (Appendix 3). The GC3 had the highest correlation with total GC and very close to the total GC value in DENV1-4, suggesting a strong mutational pressure on the third position of codons. The ENC or ENCp *v.s.* GC3 plots showed that, in addition to compositional constraint, some other factors have effect on the codon usage variations. GC2 has not important correlation with total GC in the examined genomes in the present study, implying that the constraint on this codon position is possibly due to the functional selection. A recent paper showed that the mutations on this position in the analyzed samples were mostly nonsynonymous substitutions (Behura and Severson, 2013). These results demonstrated that

both mutational and purifying selection pressures are the major forces in influencing the codon usage among DENV, consistent with some previous reports (Holmes, 2003; Jenkins and Holmes, 2003), but these factors have distinct pressure on specific nucleotide position of a codon.

The analysis of ENC showed an overall weak codon usage bias, shown in Appendix 2, where DENV2 has the highest codon bias (48.80) and DENV4 has the lowest (50.87). This result is similar to a recent report (Ma et al., 2013), indicating that the result was not affected by an increased number of samples and might represent an inherent feature of DENV. One plausible explanation could be that DENV4 is less adapted to human environment, whereas DENV2 is more adapted to humans. On the other hand, DENV2 has been associated to more aggressive diseases forms, and is generally the most prevailing serotype during outbreaks situations (Cologna et al., 2005). These could mean that codon bias of DENV2 contribute to successful infection in human cells in comparison with DENV4.

Moreover, the CA and HCbT analyses within each serotype showed similar clustering patterns for the four serotypes. The DENV strains occurring in the same continental region are more closely related, forming a cluster, indicating that viruses from a geographical group show similar codon usage bias. The Asian genomes of the four serotypes showed a wide diversity in the clusters and each of them can be further divided into more homogenous subgroups. This more diversified clustering could be the consequence of longer times of DENV evolution in Asia than in other regions. Some of Asian genomes were clustered closely to the Americans, implying an evolutionary link between Asian and American clusters. The North and South American strains tend to cluster more homogenously together with less codon usage variations, corresponding to the previous observation that a limited nucleotide

diversity exists in American DENV strains (Behura and Severson, 2013; Rivera-Osorio et al., 2011). As the DENV in North and South America came from Asia, the homogenous cluster in North and South American population could indicate a simple event of introduction from Asia, then spreading over this continent with much less adaptation time than in Asia, as the consequence of founder effect.

The sequences isolated from mosquito and monkey genomes in the CA were also grouped with human strains from the same geographic origin, indicating that sylvatic DENV changes in adaptation on codon usage in a similar way to endemic human DENV, as indicated by the study on nucleotide sequences (Vasilakis et al., 2007a). On the other hand, Zhou et al reported no link of geographic origin on the codon usage of DENV (Zhou et al., 2013). Behura and Severson found that the silent sites are favoring the geographical diversification (Behura and Severson, 2013). Our study showed that not only GC3, but also GC1 and GC2 have a good correlation with Axis major variation, depending on the serotype, suggesting that all the codon sites are related to clustering of geographical strains. Thus, the present study demonstrated the strong influence of geographic origin of DENV on shaping codon usage patterns. The discrepancy in results from studies may be due to the magnitude of samples for analysis.

The clustering groups based on the codon usage datasets or phylogenetic tree on nucleotide sequence dataset showed the similar clustering results. This observation indicates the influence of the species evolution of DENV at the level of codon usage.

8.3. Aminoacid signatures.

In this study, we analyzed the aminoacid polymorphism for each DENV protein. We found a high frequency of aminoacid polymorphism in DENV 2, reflecting the widespread intra-serotype genetic variation, in comparison to the other ones. Interestingly, the serotype 4 showed a similar high frequency of polymorphism. In general, if more genomes sequenced, more polymorphisms could be found. However, the sequenced genomes from DENV4 are much less than other serotype's genomes and showed similar high frequency to DENV2 and much higher than DENV1 and DENV3. It means that if the same quantities of genomes are compared, DENV4 should show more frequency of polymorphisms than other serotypes. This result could indicate that DENV4 has less selection pressure in the hosts (mosquito and human). As we know, the DENV4 is much less virulent than other serotypes, which could provide the opportunity to have more. The aminoacid changes in the position 390 (N D) from the E protein of DENV2 was previously associated to the DHF in the sequences from different regions (Leitmeyer et al., 1999). We found the N aminoacid predominates over D aminoacid in the geographic populations except the samples from oceanic regions, it could indicate that this change is not related to the development of DHF.

8.4. Epitopes.

The process and causes implicated in protective *v.s.* pathological immunity could be deciphered throughout of the study of epitope reactivity in different disease states. Such information may help distinguish the immune reactivity in relation with severe disease or uncomplicated disease resolution, and may assist the understanding of host- and pathogen-mediated immune response.

The core process for DENV immunity and/or immunopathology may be related to the determinant recognition receptor that leads to protection or production of sub-neutralizing antibody (Vaughan et al., 2010), thus the study of these processes at the epitope level is important. We retrieved epitope data associated with defined disease states, dengue fever (DF), dengue hemorrhagic fever. The majority of epitopes reported in IEDB defined for DF were B-cell epitopes, whereas for DHF were T-cell epitopes.

The reason for explaining this phenomenon is that neutralizing antibodies play a critical role in the normal course of DF disease, and that T-cell dysfunction is believed to be endorsed for the immunopathology leading to DHF. Some studies that support associations for protection from severe disease are the identification of neutralizing antibodies directed at the E protein, preM/M, and NS1 (Kurane et al., 1991; Lobigs et al., 1994; Roehrig et al., 1998; Rothman, 2004). On the other hand, neutralizing antibody and virus-specific T cells to these same antigens were also associated with enhancement of disease (Morens and Halstead, 1990). The mechanisms for disease exacerbation were put forward to include antibody-mediated enhanced viral uptake and or T-cell (Vaughan et al., 2010). Further we mapped the majority of epitopes in a large number of DENV proteins from different continental regions. Most of epitopes were well conserved; it could mean that the pathogenic potential is an inherent feature that exists across over the world isolates.

9. CONCLUSIONS

We have analyzed the DENV genomes from several aspects summarizing the following concluding remarks.

The phylogenetic analysis based on the E protein showed no introduction of new serotypes to the country.

The Mexican sequences remains genetically stable. The codon usage of DENV genomes was analyzed on a large scale and demonstrated that both mutational and purifying selection pressures have important contribution to the codon usage; however, these factors have distinct pressure on specific codon nucleotide positions.

The codon usage patterns of DENV genomes showed apparent geographic feature.

The aminoacid patterns are also a genomic geographic signature.

The epitopes that could be related to DENV disease pathogenesis were identified.

DENV4 showed more diverse than other serotypes.

10. RECOMENDATIONS

The genetic or genomic analysis for DENV Mexican strains should be generally implemented in the health system of the country to identify the introduction of new strains or dynamics of genetic viral population, which should help to take decisions for the control. Codon usage and geographic signatures could be employed as complementary measures for the monitoring of DENV.

The analysis of epitope reactivity of DENV related to clinical disease should be a main priority for upcoming experimental studies.

12. REFERENCES

Acosta-Bas, C., Gómez-Cordero, I., 2005. Biología y métodos diagnósticos del dengue. Rev Biomed 16, 113-137.

Alexander Diaz-Quijano, F., Arali Martinez-Vega, R., Elvira Ocazionez, R., Angel Villar-Centeno, L., 2006. [Evaluation of IgM determination in acute serum for the diagnosis of dengue in an endemic area]. Enfermedades infecciosas y microbiologia clinica 24, 90-92.

Alvarez, D.E., Lodeiro, M.F., Luduena, S.J., Pietrasanta, L.I., Gamarnik, A.V., 2005. Long-range RNA-RNA interactions circularize the dengue virus genome. J Virol 79, 6631-6643.

Avirutnan, P., Punyadee, N., Noisakran, S., Komoltri, C., Thiemmeca, S., Auethavornanan, K., Jairungsri, A., Kanlaya, R., Tangthawornchaikul, N., Puttikhunt, C., Pattanakitsakul, S.N., Yenchitsomanus, P.T., Mongkolsapaya, J., Kasinrerk, W., Sittisombut, N., Husmann, M., Blettner, M., Vasanawathana, S., Bhakdi, S., Malasit, P., 2006. Vascular leakage in severe dengue virus infections: a potential role for the nonstructural viral protein NS1 and complement. The Journal of infectious diseases 193, 1078-1088.

Baldauf, S.L., 2003. Phylogeny for the faint of heart: a tutorial. Trends Genet 19, 345-351.

Barker, W.C., Mazumder, R., Vasudevan, S., Sagripanti, J.L., Wu, C.H., 2009. Sequence signatures in envelope protein may determine whether flaviviruses produce hemorrhagic or encephalitic syndromes. Virus genes 39, 1-9.

Beaumier, C.M., Rothman, A.L., 2009. Cross-reactive memory CD4+ T cells alter the CD8+ T-cell response to heterologous secondary dengue virus infections in mice in a sequence-specific manner. Viral immunology 22, 215-219.

Behura, S.K., Severson, D.W., 2013. Nucleotide substitutions in dengue virus serotypes from Asian and American countries: insights into intracodon recombination and purifying selection. BMC Microbiol 13, 37.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R., Simmons, C.P., Scott, T.W., Farrar, J.J., Hay, S.I., 2013. The global distribution and burden of dengue. Nature 496, 504-507.

Burke, D.S., Kliks, S., 2006. Antibody-dependent enhancement in dengue virus infections. The Journal of infectious diseases 193, 601-603; author reply 603-604.

Carey, D.E., 1971. Chikungunya and dengue: a case of mistaken identity? Journal of the history of medicine and allied sciences 26, 243-262.

Carrillo-Valenzo, E., Danis-Lozano, R., Velasco-Hernandez, J.X., Sanchez-Burgos, G., Alpuche, C., Lopez, I., Rosales, C., Baronti, C., de Lamballerie, X., Holmes, E.C., Ramos-Castaneda, J., 2010. Evolution of dengue virus in Mexico is characterized by frequent lineage replacement. Archives of virology 155, 1401-1412.

Carvalho, S.E., Martin, D.P., Oliveira, L.M., Ribeiro, B.M., Nagata, T., 2010. Comparative analysis of American Dengue virus type 1 full-genome sequences. Virus genes 40, 60-66.

Clyde, K., Kyle, J.L., Harris, E., 2006. Recent advances in deciphering viral and host determinants of dengue virus replication and pathogenesis. J Virol 80, 11418-11431.

Cologna, R., Armstrong, P.M., Rico-Hesse, R., 2005. Selection for virulent dengue viruses occurs in humans and mosquitoes. J Virol 79, 853-859.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome research 14, 1188-1190.

Chambers, T.J., Hahn, C.S., Galler, R., Rice, C.M., 1990. Flavivirus genome organization, expression, and replication. Annual review of microbiology 44, 649-688.

Charif, D., Lobry, J.R., 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis Structural

Approaches to Sequence Evolution, in: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (Eds.). Springer Berlin Heidelberg, pp. 207-232.

Chaturvedi, U.C., Agarwal, R., Elbishbishi, E.A., Mustafa, A.S., 2000. Cytokine cascade in dengue hemorrhagic fever: implications for pathogenesis. FEMS immunology and medical microbiology 28, 183-188.

Chen, Y., Maguire, T., Hileman, R.E., Fromm, J.R., Esko, J.D., Linhardt, R.J., Marks, R.M., 1997. Dengue virus infectivity depends on envelope protein binding to target cell heparan sulfate. Nature medicine 3, 866-871.

Chiu, W.W., Kinney, R.M., Dreher, T.W., 2005. Control of translation by the 5'- and 3'-terminal regions of the dengue virus genome. J Virol 79, 8303-8315.

Christenbury, J.G., Aw, P.P., Ong, S.H., Schreiber, M.J., Chow, A., Gubler, D.J., Vasudevan, S.G., Ooi, E.E., Hibberd, M.L., 2010. A method for full genome sequencing of all four serotypes of the dengue virus. Journal of virological methods 169, 202-206.

D'Arcy, A., Chaillet, M., Schiering, N., Villard, F., Lim, S.P., Lefeuvre, P., Erbel, P., 2006. Purification and crystallization of dengue and West Nile virus NS2B-NS3 complexes. Acta crystallographica. Section F, Structural biology and crystallization communications 62, 157-162.

Delgado, I., Vazquez, S., Bravo, J.R., Gúzman, M.J., 2002. Predicción del serotipo del virus del dengue mediante la respuesta de anticuerpos IgM. Revista Cubana de Medicina Tropical. 54, 111–115.

Diaz, F.J., Black, W.C.t., Farfan-Ale, J.A., Lorono-Pino, M.A., Olson, K.E., Beaty, B.J., 2006. Dengue virus circulation and evolution in Mexico: a phylogenetic perspective. Archives of medical research 37, 760-773.

Dray, S., Dufor, A.B., 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software 22, 1-20.

Durán, C.A., Lanza, T.M., Plata, J.A., 2010. Fisiopatología y diagnóstico del dengue. Revista Médica de Honduras 78, 1–6.

Erbel, P., Schiering, N., D'Arcy, A., Renatus, M., Kroemer, M., Lim, S.P., Yin, Z., Keller, T.H., Vasudevan, S.G., Hommel, U., 2006. Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. Nature structural & molecular biology 13, 372-373.

Etherington, G.J., Dicks, J., Roberts, I.N., 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics 21, 278-281.

Filomatori, C.V., Lodeiro, M.F., Alvarez, D.E., Samsa, M.M., Pietrasanta, L., Gamarnik, A.V., 2006. A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. Genes & development 20, 2238-2249.

Gomez-Dantes, H., Willoquet, J.R., 2009. Dengue in the Americas: challenges for prevention and control. Cadernos de saude publica 25 Suppl 1, S19-31.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221-224.

Greenacre, M., 2007. Correspondence Analysis in Practice. Chapman & Hall/CRC, London second ed.

Gubler, D.J., 1997. Dengue and dengue hemorrhagic fever: its history and resurgence as a global public health problem. . In: Gubler, D.J., Kuno, G. (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI Publishing, Oxon, , 1–22.

Gubler, D.J., 2002. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. Trends in microbiology 10, 100-103.

Guha-Sapir, D., Schimmer, B., 2005. Dengue fever: new paradigms for a changing epidemiology. Emerging themes in epidemiology 2, 1.

Gurugama, P., Garg, P., Perera, J., Wijewickrama, A., Seneviratne, S.L., 2010. Dengue viral infections. Indian journal of dermatology 55, 68-78.

Guzmán, M.D., Vázquez, S., 2002. Apuntes sobre el diagnóstico de laboratorio del virus dengue. Revista Cubana de Medicina Tropical, 54, 180–188.

Halstead, S.B., 2003. Neutralization and antibody-dependent enhancement of dengue viruses. Advances in virus research 60, 421-467.

Hardison, R.C., 2003. Comparative genomics. PLoS biology 1, E58.

Henchal, E.A., Putnak, J.R., 1990. The dengue viruses. Clinical microbiology reviews 3, 376-396.

Holmes, E.C., 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol 77, 11296-11298.

Jelinek, T., 2000. Dengue fever in international travelers. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 31, 144-147.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Research 92, 1-7.

Jenkins, G.M., Pagel, M., Gould, E.A., de, A.Z.P.M., Holmes, E.C., 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J Mol Evol 52, 383-390.

Jessie, K., Fong, M.Y., Devi, S., Lam, S.K., Wong, K.T., 2004. Localization of dengue virus in naturally infected human tissues, by immunohistochemistry and in situ hybridization. The Journal of infectious diseases 189, 1411-1418.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780.

Khumthong, R., Angsuthanasombat, C., Panyim, S., Katzenmeier, G., 2002. In vitro determination of dengue virus type 2 NS2B-NS3 protease activity with fluorescent peptide substrates. Journal of biochemistry and molecular biology 35, 206-212.

Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., Lundegaard, C., Sette, A., Lund, O., Bourne, P.E., Nielsen, M., Peters, B., 2012. Immune epitope database analysis resource. Nucleic Acids Res 40, W525-530.

Kliks, S.C., Nimmanitya, S., Nisalak, A., Burke, D.S., 1988. Evidence that maternal dengue antibodies are important in the development of dengue hemorrhagic fever in infants. The American journal of tropical medicine and hygiene 38, 411-419.

Kuhn, R.J., Zhang, W., Rossmann, M.G., Pletnev, S.V., Corver, J., Lenches, E., Jones, C.T., Mukhopadhyay, S., Chipman, P.R., Strauss, E.G., Baker, T.S., Strauss, J.H., 2002. Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. Cell 108, 717-725.

Kuno, G., 1995. Review of the factors modulating dengue transmission. Epidemiologic reviews 17, 321-335.

Kurane, I., Brinton, M.A., Samson, A.L., Ennis, F.A., 1991. Dengue virus-specific, human CD4+ CD8- cytotoxic T-cell clones: multiple patterns of virus cross-reactivity recognized by NS3-specific T-cell clones. J Virol 65, 1823-1828.

Lanciotti, R.S., Calisher, C.H., Gubler, D.J., Chang, G.J., Vorndam, A.V., 1992. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. J Clin Microbiol 30, 545-551.

Leitmeyer, K.C., Vaughn, D.W., Watts, D.M., Salas, R., Villalobos, I., de, C., Ramos, C., Rico-Hesse, R., 1999. Dengue virus structural differences that correlate with pathogenesis. J Virol 73, 4738-4747.

Libraty, D.H., Pichyangkul, S., Ajariyakhajorn, C., Endy, T.P., Ennis, F.A., 2001. Human dendritic cells are activated by dengue virus infection: enhancement by gamma interferon and implications for disease pathogenesis. J Virol 75, 3501-3508.

Libraty, D.H., Young, P.R., Pickering, D., Endy, T.P., Kalayanarooj, S., Green, S., Vaughn, D.W., Nisalak, A., Ennis, F.A., Rothman, A.L., 2002. High circulating levels of the dengue virus nonstructural protein NS1 early in dengue illness correlate with the development of dengue hemorrhagic fever. The Journal of infectious diseases 186, 1165-1168.

Lobigs, M., Arthur, C.E., Mullbacher, A., Blanden, R.V., 1994. The flavivirus nonstructural protein NS3 is a dominant source of cytotoxic T cell peptide determinants. Virology 202, 195-201.

Ma, J.-J., Zhao, F., Zhang, J., Zhou, J.-H., Ma, L.-N., Ding, Y.-Z., Chen, H.-T., Gu, Y.-X., Yong-Sheng Liu, 2013. Analysis of Synonymous Codon Usage in Dengue Viruses. Journal of Animal and Veterinary Advances 12, 88-98.

Ma, L., Jones, C.T., Groesch, T.D., Kuhn, R.J., Post, C.B., 2004. Solution structure of dengue virus capsid protein reveals another fold. Proceedings of the National Academy of Sciences of the United States of America 101, 3414-3419.

Miller, W., Makova, K.D., Nekrutenko, A., Hardison, R.C., 2004. Comparative genomics. Annual review of genomics and human genetics 5, 15-56.

Ming-Wei, S., Chu, W.C., Yuan, H.S., 2007. Distinguish Dengue Virus Serotypes via Codon Usage Patterns, Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, pp. 1328-1330.

Modis, Y., Ogata, S., Clements, D., Harrison, S.C., 2003. A ligand-binding pocket in the dengue virus envelope glycoprotein. Proceedings of the National Academy of Sciences of the United States of America 100, 6986-6991.

Montes, M.T., 2001. Actualización en dengue, parte 1. Revista Sociedad Venezolana de Microbiologia 21, 1-12.

Morens, D.M., Halstead, S.B., 1990. Measurement of antibody-dependent infection enhancement of four dengue virus serotypes by monoclonal and polyclonal antibodies. J Gen Virol 71 (Pt 12), 2909-2914.

Murillo-Llanes, J., Soto-Valenzuela, H., Flores-Flores, P., Pereza-Garay, F., 2007. Caracterización clínica y epidemiológica del dengue. Rev Med Inst Mex Seguro Soc 45, 485-491.

Noisakran, S., Perng, G.C., 2008. Alternate hypothesis on the pathogenesis of dengue hemorrhagic fever (DHF)/dengue shock syndrome (DSS) in dengue virus infection. Experimental biology and medicine 233, 401-408.

Novembre, J.A., 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19, 1390-1394.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Peden, J.F., 1999. Analysis of Codon Usage. Avalaible at: http://codonw.sourceforge.net/.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.

R-Development-Core-Team., 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rajapakse, S., Rodrigo, C., Maduranga, S., Rajapakse, A.C., 2014. Corticosteroids in the treatment of dengue shock syndrome. Infect Drug Resist 7, 137-143.

Resch, W., Zaslavsky, L., Kiryutin, B., Rozanov, M., Bao, Y., Tatusova, T.A., 2009. Virus variation resources at the National Center for Biotechnology Information: dengue virus. BMC Microbiology 9, 65.

Reyes-Del Valle, J., Chavez-Salinas, S., Medina, F., Del Angel, R.M., 2005. Heat shock protein 90 and heat shock protein 70 are components of dengue virus receptor complex in human cells. J Virol 79, 4557-4567.

Rivera-Osorio, P., Vaughan, G., Ramirez-Gonzalez, J.E., Fonseca-Coronado, S., Ruiz-Tovar, K., Cruz-Rivera, M.Y., Ruiz-Pacheco, J.A., Vazquez-Pichardo, M., Carpio-Pedroza, J.C., Cazares, F., Escobar-Gutierrez, A., 2011. Molecular epidemiology of autochthonous dengue virus strains circulating in Mexico. J Clin Microbiol 49, 3370-3374.

Rodenhuis-Zybert, I.A., Wilschut, J., Smit, J.M., 2010. Dengue virus life cycle: viral and host factors modulating infectivity. Cellular and molecular life sciences : CMLS 67, 2773-2786.

Roehrig, J.T., Bolin, R.A., Kelly, R.G., 1998. Monoclonal antibody mapping of the envelope glycoprotein of the dengue 2 virus, Jamaica. Virology 246, 317-328.

Rothman, A.L., 2004. Dengue: defining protective versus pathologic immunity. The Journal of clinical investigation 113, 946-951.

Rush, B., 1789. An account of the bilious remitting fever, as it appeared in Philadelphia in the summer and autumn of the year 1789. Medical Inquiries and Observations., 104-126.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15, 1281-1295.

Smith, C.E., 1956. The history of dengue in tropical Asia and its probable relationship to the mosquito Aedes aegypti. The Journal of tropical medicine and hygiene 59, 243-251.

Suzuki, H., Brown, C.J., Forney, L.J., Top, E.M., 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res 15, 357-365.

Tassaneetrithep, B., Burgess, T.H., Granelli-Piperno, A., Trumpfheller, C., Finke, J., Sun, W., Eller, M.A., Pattanapanyasat, K., Sarasombath, S., Birx, D.L., Steinman, R.M., Schlesinger, S., Marovich, M.A., 2003. DC-SIGN (CD209) mediates dengue virus infection of human dendritic cells. The Journal of experimental medicine 197, 823-829.

Taylor, R., 1990. Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography 6, 35-39.

Thurner, C., Witwer, C., Hofacker, I.L., Stadler, P.F., 2004. Conserved RNA secondary structures in Flaviviridae genomes. J Gen Virol 85, 1113-1124.

Tolou, H.J., Couissinier-Paris, P., Durand, J.P., Mercier, V., de Pina, J.J., de Micco, P., Billoir, F., Charrel, R.N., 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, 1283-1290.

Tomlinson, S.M., Malmstrom, R.D., Watowich, S.J., 2009. New approaches to structure-based discovery of dengue protease inhibitors. Infectious disorders drug targets 9, 327-343.

Twiddy, S.S., Woelk, C.H., Holmes, E.C., 2002. Phylogenetic evidence for adaptive evolution of dengue viruses in nature. J Gen Virol 83, 1679-1689.

Vasilakis, N., Holmes, E.C., Fokam, E.B., Faye, O., Diallo, M., Sall, A.A., Weaver, S.C., 2007a. Evolutionary Processes among Sylvatic Dengue Type 2 Viruses. Journal of Virology 81, 9591-9595.

Vasilakis, N., Shell, E.J., Fokam, E.B., Mason, P.W., Hanley, K.A., Estes, D.M., Weaver, S.C., 2007b. Potential of ancestral sylvatic dengue-2 viruses to re-emerge. Virology 358, 402-412.

Vaughan, K., Greenbaum, J., Blythe, M., Peters, B., Sette, A., 2010. Meta-analysis of all immune epitope data in the Flavivirus genus: inventory of current immune epitope data status in the context of virus immunity and immunopathology. Viral immunology 23, 259-284.

Vaughn, D.W., Green, S., Kalayanarooj, S., Innis, B.L., Nimmannitya, S., Suntayakorn, S., Endy, T.P., Raengsakulrach, B., Rothman, A.L., Ennis, F.A., Nisalak, A., 2000. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. The Journal of infectious diseases 181, 2-9.

Velandia, M.L., Castellanos, J.E., 2011. Virus del dengue: estructura y ciclo viral. Infectio 15, 33-43.

Wang, E., Ni, H., Xu, R., Barrett, A.D., Watowich, S.J., Gubler, D.J., Weaver, S.C., 2000. Evolutionary relationships of endemic/epidemic and sylvatic dengue viruses. J Virol 74, 3227-3234.

WHO, 2011. Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. WHO Library Cataloguing-in-Publication data, 1-192.

Worobey, M., Holmes, E.C., 1999. Evolutionary aspects of recombination in RNA viruses. J Gen Virol 80 (Pt 10), 2535-2543.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23-29.

Zanotto, P.M., Gould, E.A., Gao, G.F., Harvey, P.H., Holmes, E.C., 1996. Population dynamics of flaviviruses revealed by molecular phylogenies. Proceedings of the National Academy of Sciences of the United States of America 93, 548-553.

Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P.E., Bui, H.H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Lundegaard, C., Nielsen, M., Ponomarenko, J., Sette, A., Zhu, Z., Peters, B., 2008. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res 36, W513-518.

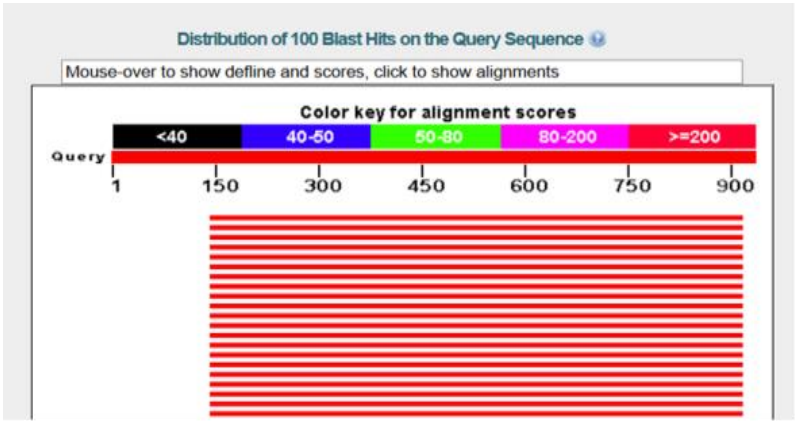
Zhang, W., Chipman, P.R., Corver, J., Johnson, P.R., Zhang, Y., Mukhopadhyay, S., Baker, T.S., Strauss, J.H., Rossmann, M.G., Kuhn, R.J., 2003. Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. Nature structural biology 10, 907-912.

Zhang, Y., Zhang, W., Ogata, S., Clements, D., Strauss, J.H., Baker, T.S., Kuhn, R.J., Rossmann, M.G., 2004. Conformational changes of the flavivirus E glycoprotein. Structure 12, 1607-1618.

Zhou, J.H., Zhang, J., Sun, D.J., Ma, Q., Chen, H.T., Ma, L.N., Ding, Y.Z., Liu, Y.S., 2013. The distribution of synonymous codon choice in the translation initiation region of dengue virus. PLoS One 8, e77239.

13. APPENDIX

Appendix 1. Blast analysis using the nucleotide sequence obtained from the tested DENV 2.



Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Dengue virus type 2 strain New Guinea C polyprotein mRNA, complete cds	1367	1367	82%	0.0	99%	AF038403.1
Dengue virus 2 isolate GU/DB019/2001 envelope glycoprotein gene, partial cds	1362	1362	82%	0.0	98%	JF804032.1
Dengue virus 2 isolate DENV-2/PG/BID-V2618/2008, complete genome	1362	1362	82%	0.0	98%	FJ906959.1
Dengue virus type 2 gene for polyprotein, partial cds, strain:New Guinea C	1362	1362	82%	0.0	98%	D00346.1
Dengue virus 2 isolate DENV-2/US/BID-V5411/2006, complete genome	1356	1356	82%	0.0	98%	JF730053.1

Download GenBank Graphics

Dengue virus type 2 strain New Guinea C polyprotein mRNA, complete cds

Sequence ID: gb|AF038403.1|AF038403 Length: 10724 Number of Matches: 1

Range 1: 938 to 1707 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
1367 bits(740)	0.0	766/777(99%)	8/777(1%)	Plus/Plus

```
Query  145   TGCGCTGCATAGGAATATCAAATAGAGACTTTGTAGAAGGGGTTTCAGGAGGAAGCTGGG  204
             ||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  938   TGCGTTGCATAGGAATATCAAATAGAGACTTTGTAGAAGGGGTTTCAGGAGGAAGCTGGG  997

Query  205   TTGACATAGTCTTAGAACATGGAAGCTGTGTGACGACGATGGCAAAAAACAAACCAACAT  264
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  998   TTGACATAGTCTTAGAACATGGAAGCTGTGTGACGACGATGGCAAAAAACAAACCAACAT  1057

Query  265   TGGATTTTGAACTGATAAAAACAGAAGCCAAACAACCTGCCACTCTAAGGAAGTACTGTA  324
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1058  TGGATTTTGAACTGATAAAAACAGAAGCCAAACAACCTGCCACTCTAAGGAAGTACTGTA  1117

Query  325   TAGAGGCAAAGCTGACCAACACAACAACAGATTCTCGCTGCCCAACACAAGGAGAACCCA  384
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1118  TAGAGGCAAAGCTGACCAACACAACAACAGATTCTCGCTGCCCAACACAAGGAGAACCCA  1177
```

Appendix 2. The nucleotide composition and the Effective Number of Codons (ENC) for DENV 1-4 genomes

	DENV1	DENV2	DENV3	DENV4
GC	0.46 ±0.001	0.46 ±0.002	0.46 ±0.001	0.47 ±0.001
GC1	0.50 ±0.003	0.49 ±0.001	0.49 ±0.001	0.49 ±0.001
GC2	0.43 ±0.001	0.43 ±0.002	0.43 ±0.001	0.43 ±0.001
GC3	0.46 ±0.004	0.45 ±0.007	0.46 ±0.003	0.48 ±0.002
ENC	49.28 ±0.192	48.80 ±0.288	49.54 ±0.131	50.87 ±0.174

Appendix 3.- Correlation matrix of nucleotide composition, ENC and ENCp.

Serotype		GC	GC1	GC2	GC3	ENC
DENV-1	GC1	-0.11				
	GC2	0.22	-0.72			
	GC3	0.72	-0.74	0.51		
	ENC	0.33	-0.33	0.19	0.46	
	ENCp	0.01	-0.17	0.1	0.13	0.89
DENV-2	GC1	0.17				
	GC2	-0.1	-0.01			
	GC3	0.93	-0.03	-0.41		
	ENC	0.26	0.38	-0.01	0.17	
	ENCp	-0.22	0.36	0.18	-0.33	0.84
DENV-3	GC1	0.63				
	GC2	-0.17	-0.37			
	GC3	0.96	0.48	-0.31		
	ENC	-0.1	-0.34	0.29	-0.07	
	ENCp	-0.36	-0.45	0.32	-0.34	0.91
DENV-4	GC1	0.53				
	GC2	-0.34	-0.42			
	GC3	0.9	0.2	-0.47		
	ENC	0.02	-0.15	-0.3	0.2	
	ENCp	-0.2	-0.16	-0.19	-0.07	0.91

Appendix 4. The identified outliers in the correspondence analysis for codon usage.

	ID	Accesion	A1 position	A2 position	Region	Country origin	Cluster ubication	Reference
D1	1	AF298808	-0.008974	-0.00013	Africa	Djibouti	Asia	[1]
	2	DQ285562	-0.071635	-0.00478	Africa	Comoros	Nth and Sth America	
	4	DQ285559	-0.059043	0.011857	Africa	Reunion	Nth and Sth America	
	964	JN903579	-0.052876	0.011935	Asia	India	Nth and Sth America	[2]
	971	JN903580	-0.05706	0.016168	Asia	India	Nth and Sth America	[2]
	972	JN903581	-0.0572	0.01333	Asia	India	Nth and Sth America	[2]
	980	AY713473	-0.053197	-0.00845	Asia	Myanmar	Nth and Sth America	
	983	AY732474	-0.065983	0.001864	Asia	Thailand	Nth and Sth America	[3]
	984	AY732476	-0.065268	0.003351	Asia	Thailand	Nth and Sth America	[3]
	1054	EU081258	-0.055036	0.014598	Asia	Singapore	Nth and Sth America	[4]
	1224	DQ672562	-0.026632	0.034354	Nth America	USA Hawaii	Oceania	[5]
	1225	DQ672561	-0.026632	0.034354	Nth America	USA Hawaii	Oceania	[5]
	1226	DQ672564	-0.00083	0.027413	Nth America	USA Hawaii	Asia	[5]
	1227	DQ672563	-0.024588	0.03204	Nth America	USA Hawaii	Oceania	[5]
	1249	EU863650	-0.02098	0.040953	Sth America	Chile	Ocenia	[6]
D2	5	EU056810	0.010148	-0.05485	Africa	Burkina	Asia	[7]
	26	JN819418	-0.03524	0.005203	Asia	Viet Nam	Nth America	
	394	GQ868600	0.043494	-0.0528	Nth America	Puerto Rico	Asia	
	395	EU056812	0.051944	-0.04976	Nth America	Puerto Rico	Asia	
	396	GQ868588	0.040364	-0.05374	Nth America	Mexico	Asia	
	397	GQ868589	0.040364	-0.05374	Nth America	Mexico	Asia	
	398	FJ898449	0.04164	-0.03932	Nth America	Honduras	Asia	
	432	GQ868590	0.027059	-0.05032	Nth America	Mexico Sonora	Asia	
	658	JF730053	0.037205	-0.01011	Nth America	USA California mouse adapted	Asia	[8]
	784	JF730050	0.04007	-0.01017	Nth America	USA California mouse adapted	Asia	
	799	HQ541798	0.035418	-0.01205	Nth America	USA California mouse adapted	Asia	

	803	EU854293	0.041849	-0.01095	Sth America	Colombia	Asia	
	804	AY702040	0.045401	-0.02918	Sth America	Colombia	Asia	[9]
	805	GQ868592	0.045115	-0.02916	Sth America	Colombia	Asia	
	817	EU056811	0.048425	-0.04921	Sth America	Peru	Asia	[7]
	894	FJ390389	0.041673	-0.01184	Oceania	Papua New Guinea	Asia	
	914	HM488257	0.018495	-0.05251	Oceania	Guam Micronesia	Asia	
	915	FJ906959	0.037216	-0.01125	Oceania	Papua New Guinea	Asia	
	927	EF105379	0.048482	-0.05199	Asia	Malaysia	Asia Monkey	[10]
D3	1	FJ882575	0.004436	0.003597	Africa	Mozambique	Asia	
	483	EF629370	-0.03753	-0.01011	Sth America	Brazil	Asia	
	624	JN697379	-0.03552	-0.01032	Sth America	Brazil	Asia	
	668	JN406514	-0.05875	0.024313	Oceania	Australia	Asia	[11]
	669	JN406515	-0.05988	-0.07313	Oceania	Australia	Asia	[11]
D4	7	FJ196849	-0.01392	0.00423	Asia	China	Near to America	
	112	JQ513345	-0.04373	-0.06771	Sth America	Brazil	Asia	

References included on the table

1.-H. J. Tolou, P. Couissinier-Paris, J. P. Durand, V. Mercier, J. J. de Pina, P. de Micco, F. Billoir, R. N. Charrel,X. de Lamballerie, "Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences," J Gen Virol, vol. 82, no. Pt 6, pp. 1283-90, 2001.

2.-M. Anoop, A. J. Mathew, B. Jayakumar, A. Issac, S. Nair, R. Abraham, M. G. Anupriya,E. Sreekumar, "Complete genome sequencing and evolutionary analysis of dengue virus serotype 1 isolates from an outbreak in Kerala, South India," Virus genes, vol. 45, no. 1, pp. 1-13, 2012.

3.-C. Zhang, M. P. Mammen, Jr., P. Chinnawirotpisan, C. Klungthong, P. Rodpradit, P. Monkongdee, S. Nimmannitya, S. Kalayanarooj,E. C. Holmes, "Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence," J Virol, vol. 79, no. 24, pp. 15123-30, 2005.

4.-M. J. Schreiber, E. C. Holmes, S. H. Ong, H. S. Soh, W. Liu, L. Tanner, P. P. Aw, H. C. Tan, L. C. Ng, Y. S. Leo, J. G. Low, A. Ong, E. E. Ooi, S. G. Vasudevan,M. L. Hibberd, "Genomic epidemiology of a dengue virus epidemic in urban Singapore," J Virol, vol. 83, no. 9, pp. 4163-73, 2009.

5.-A. Imrie, C. Roche, Z. Zhao, S. Bennett, M. Laille, P. Effler,V. M. Cao-Lormeau, "Homology of complete genome sequences for dengue virus type-1, from dengue-fever- and dengue-haemorrhagic-fever-associated epidemics in Hawaii and French Polynesia," Annals of tropical medicine and parasitology, vol. 104, no. 3, pp. 225-35, 2010.

6.-C. Caceres, V. Yung, P. Araya, J. Tognarelli, E. Villagra, L. Vera,J. Fernandez, "Complete nucleotide sequence analysis of a Dengue-1 virus isolated on Easter Island, Chile," Archives of virology, vol. 153, no. 10, pp. 1967-70, 2008.

7.-N. Vasilakis, E. B. Fokam, C. T. Hanson, E. Weinberg, A. A. Sall, S. S. Whitehead, K. A. Hanley,S. C. Weaver, "Genetic and phenotypic characterization of sylvatic dengue virus type 2 strains," Virology, vol. 377, no. 2, pp. 296-307, 2008.

8.-S. Orozco, M. A. Schmid, P. Parameswaran, R. Lachica, M. R. Henn, R. Beatty,E. Harris, "Characterization of a model of lethal dengue virus 2 infection in C57BL/6 mice deficient in the alpha/beta interferon receptor," J Gen Virol, vol. 93, no. Pt 10, pp. 2152-7, 2012.

9.-R. Rodriguez-Roche, M. Alvarez, T. Gritsun, S. Halstead, G. Kouri, E. A. Gould,M. G. Guzman, "Virus evolution during a severe dengue epidemic in Cuba, 1997," Virology, vol. 334, no. 2, pp. 154-9, 2005.

10.-N. Vasilakis, E. C. Holmes, E. B. Fokam, O. Faye, M. Diallo, A. A. Sall,S. C. Weaver, "Evolutionary Processes among Sylvatic Dengue Type 2 Viruses," Journal of Virology, vol. 81, no. 17, pp. 9591-9595, 2007.

11.-S. A. Ritchie, A. T. Pyke, S. Hall-Mendelin, A. Day, C. N. Mores, R. C. Christofferson, D. J. Gubler, S. N. Bennett,A. F. van den Hurk, "An explosive epidemic of DENV-3 in Cairns, Australia," PLoS One, vol. 8, no. 7, pp. e68137, 2013.

Appendix 5. Paper result from this study

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 851425, 9 pages
http://dx.doi.org/10.1155/2014/851425

Research Article

Large-Scale Genomic Analysis of Codon Usage in Dengue Virus and Evaluation of Its Phylogenetic Dependence

Edgar E. Lara-Ramírez,[1] Ma Isabel Salazar,[2] María de Jesús López-López,[1] Juan Santiago Salas-Benito,[3] Alejandro Sánchez-Varela,[1] and Xianwu Guo[1]

[1] *Laboratory of Molecular Biomedicine, Center of Biotechnology on Genomics, National Polytechnic Institute, Colonia Narciso Mendoza, 88710 Reynosa, TAMPS, Mexico*
[2] *Laboratory for Cellular Immunology and Immunopathogenesis, Department of Immunology, National School for Biological Sciences (ENCB), National Polytechnic Institute, 11340 New Mexico, DF, Mexico*
[3] *Laboratory for Biomedicine, Department of Virology, National School of Medicine and Homeopathy, National Polytechnic Institute, 11340 New Mexico, DF, Mexico*

Correspondence should be addressed to Xianwu Guo; gxianwu@yahoo.com

Received 25 February 2014; Revised 5 June 2014; Accepted 11 June 2014; Published 17 July 2014

Academic Editor: Sankar Subramanian

The increasing number of dengue virus (DENV) genome sequences available allows identifying the contributing factors to DENV evolution. In the present study, the codon usage in serotypes 1–4 (DENV1–4) has been explored for 3047 sequenced genomes using different statistics methods. The correlation analysis of total GC content (GC) with GC content at the three nucleotide positions of codons (GC1, GC2, and GC3) as well as the effective number of codons (ENC, ENCp) versus GC3 plots revealed mutational bias and purifying selection pressures as the major forces influencing the codon usage, but with distinct pressure on specific nucleotide position in the codon. The correspondence analysis (CA) and clustering analysis on relative synonymous codon usage (RSCU) within each serotype showed similar clustering patterns to the phylogenetic analysis of nucleotide sequences for DENV1–4. These clustering patterns are strongly related to the virus geographic origin. The phylogenetic dependence analysis also suggests that stabilizing selection acts on the codon usage bias. Our analysis of a large scale reveals new feature on DENV genomic evolution.

1. Introduction

Dengue virus (DENV) is a positive strand RNA virus that belongs to the Flaviviridae family [1]. Its genome is approximately 11 kb long with an uninterrupted open reading frame (ORF) that encodes a polyprotein. DENV commonly exists as four (DENV1–4) distinct but genetically related serotypes. A new serotype (DENV5) has been recently described [2]. DENV exists in either sylvatic or human transmission cycles [3], which are most prevalent in tropical and subtropical areas, where ecoepidemiologic conditions contribute to sustaining the virus in nature. According to the World Health Organization, ≈2.5 billion people living in >100 countries are at risk of being infected by one or more of the DENV serotypes [4]. DENV are the cause of dengue fever and the more complicated forms of diseases, dengue haemorrhagic fever and dengue shock syndrome. At the present, there is no effective vaccine to prevent dengue diseases and no drug for specific therapy.

The degeneracy is an intrinsic characteristic of genetic code and enables different codons to encode for a given amino acid. However, the choice of synonymous codons is not random for a species; therefore, codon usage varies among species [5]. Some factors seem to influence the codon usage; for example, mutational bias has been attributed as the major determinant of codon usage variation among RNA viruses [6]. In addition, the codon usage deviations are the evolutionary consequence of an organism [5] and the result of adaptive interaction between pathogenic viruses and their hosts [7]. Thus, it has been proposed that codon usage

is useful to discern the evolutionary relationships between species [8] and the patterns of codon variation may also shed some light on fundamental questions on basic biology.

The analyses of codon usage in DENV have been previously studied in the context of genus *flavivirus* [7, 9, 10], RNA type viruses [6, 11], or DENV genomic comparisons [12–15]. These studies have provided some valuable information; however, only a limited number of genomes were employed for their analysis. The increasing number of genome sequences reported from all over the world could thus help to reveal how DENV genomes diverge and what the principal contributing factors for their evolution are. Here, the genome-wide codon usage patterns were analyzed for 3047 full-length genomes of DENV1–4. In addition, we applied two methods to assess the phylogenetic dependence of codon usage to unravel novel evolutionary features of DENV.

2. Materials and Methods

2.1. Genome Sequences. The whole genome sequences of 3047 DENV1–4 were downloaded from the NCBI DENV resource at http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/nph-select.cgi?taxid=12637. This website provided DENV information that includes sample sequence, location, and serotype [16]. Four datasets that correspond to each one of the four serotypes were established. They included 1336 genomes for DENV1, 927 genomes for DENV2, 670 genomes for DENV3, and 114 genomes for DENV4. The coding sequences of genomes were collected in a dataset for each serotype orderly according to their geographic regions of isolation as Africa, Asia, North America, Oceania, and South America and for the samples from the same continent along with the order of host sources as human, mosquito, monkey, and unknown host. A number was then assigned to each genome in each dataset, which facilitates the subsequent analyses. The accession numbers as well as the assigned numbers for corresponding genomes in the present study are provided in an excel spreadsheet in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/851425.

2.2. Nucleotide Compositions and Codon Usage Bias. The total GC% (GC) and GC% at the 1st (GC1), 2nd (GC2), and 3rd (GC3) codon positions of coding sequences for each DENV genome sequence were calculated in order to show the impact of selection on codon usage of DENV. The total GC content was calculated with the following equation:

$$\mathrm{GC} = \frac{(\mathrm{G} + \mathrm{C})}{(\mathrm{A} + \mathrm{T} + \mathrm{G} + \mathrm{C})}, \tag{1}$$

where the G, C, A, and T are the number of nucleotides in the genome. For the calculation of GC at the three codon positions we used the following equation:

$$\mathrm{GC}n = \frac{\mathrm{G}n + \mathrm{C}n}{(L/3)}, \tag{2}$$

where Gn, Cn are the number of guanines and cytosines at the nth (1, 2, or 3) position of the codon and L is the length of the genome.

Relative synonymous codon usage (RSCU) [17] was estimated as a proportion of the observed occurrence of codons to the expected occurrence when all codons for the same amino acid are equally used. The RSCU was calculated with the following equation:

$$\mathrm{RSCU} = \frac{X_{ij}}{\sum_j^{ni} X_{ij}} n_i, \tag{3}$$

where X_{ij} is the observed number of the ith codon for the jth amino acid which has n_i kinds of synonymous codons. It was measured for 59 codons except Met, Trp, and the three stop codons for each genome tested in this study. Effective number of codons (ENC) is a parameter to reveal the number of equally used codons that could yield the observed codon usage bias in a gene or a genome [18]. ENC was calculated with the following equation:

$$\mathrm{ENC} = 2 + \frac{9}{\overline{F}2} + \frac{1}{\overline{F}3} + \frac{5}{\overline{F}4} + \frac{3}{\overline{F}6}, \tag{4}$$

where $\overline{F}k$ ($k = 2, 3, 4, 6$) is the mean of $\overline{F}k$ values for the k-fold degenerate amino acids. ENC's values range from 20, the strongest bias, to 61, no bias. Because the genomes of DENV have unique uninterrupted polyprotein ORF, we applied ENC to quantify the level of codon usage bias on genome level in the present study. ENC prime (ENCp) was also used to quantify the codon bias taking into account the nucleotide background of the genomes [19]. The GC at three codon positions and RSCU were calculated with package seqinr [20] for R [21], and ENC and ENCp were calculated with the software Codonw and the software ENC prime, respectively [19, 22].

2.3. Correspondence Analysis. Correspondence analysis (CA) is an effective method to show the relationship among multiple categorical variables by a statistical procedure. The unique condition is to have a nonnegative data ordered in a two-way table for analysis. It is much better if the table consists of large enough dataset and homogenous variables [23]. Our RSCU dataset forms a table that should meet well the CA conditions. The RSCU table was read and formatted as data.frame in order to perform the CA with the function "*dudi.coa*" using the ADE-4 package [24] in R. In the results obtained, each genome was represented as 59-orthogonal axes, and each axis corresponds to one of 59 codons. Thus, the results of CA show how much DENV genomes are correlated to the level of codon usage variation patterns. The advantage of CA is that the results can be depicted as a map, in which each row and each column are represented as a point, which facilitates the understanding of the relation of codon usage bias among the genomes.

2.4. Evaluation of Influencing Evolutionary Factors of DENV Codon Usage. Correlation analysis of GC1, GC2, GC3, GC, ENC, and ENCp values and the selected axis of variation of each DENV1–4 dataset was performed, using Pearson's rank correlation method. For better explanation of the correlation results, only the coefficient ≥ 0.70 was considered as strong correlation [25, 26]. As regards the evaluation of correlation

coefficients, the null hypothesis of no correlation between the variables was tested at significance level of $P = 0.01$.

2.5. Hierarchical Clustering Based on Codon Usage. A distance matrix that accounts for differences in RSCUs for DENV genomes was constructed with the function "*dist*" and the Euclidean distance method by the software R. The matrix obtained was then used to aggregate the RSCU values of each genome sequence into hierarchical clusters of similar codon patterns with the function "*hclust*" and the Ward method by the software R. The *hclust* objects produced were then transformed to phylo objects for plotting the final trees with the ape [27] and pyloch packages for R.

2.6. Alignments, Phylogenetic Trees, and Recombination Analysis. A phylogenetic analysis was also performed, based on the nucleotide sequences of genomes, to compare the result with that of clustering analysis based on the codon usage. The software MAFFT was used to align the whole DENV genomes of coding regions [28]. We used the default "*—auto*" function to run the alignments on MAFFT. The FastTree [29] software was used to construct approximately maximum-likelihood phylogenetic trees for each of DENV1–4 from alignments data. FastTree software can handle large alignments in a practical amount of time and memory. The generalized time-reversible (GTR) model was used for phylogenetic tree construction. To estimate the local support values of each split in the tree, the Shimodaira-Hasegawa test was used. The *Newick* tree files generated were used with the ape and pyloch packages for R to plot the phylogenetic trees. As the recombination has also impact on the evolution of DENV [30], we also tested this pattern using the software Recombination Analysis Tool (RAT) [31] for each DENV dataset.

2.7. Evaluation of Phylogenetic Dependence of Codon Usage. Phylogenetic dependence is a frequently employed test to evaluate the correlation of phenotypical traits with phylogenetic tree [32]. Such analysis was recently applied for codon usage bias in mosquitos [33]. In our study, two measures [34] (Abouheif's Cmean and Blomberg's K) have been applied to evaluate the dependence of codon usage values with the inferred phylogenetic tree of DENV1–4. To estimate Abouheif's Cmean, we firstly constructed phylo4d objects which contain the combined DENV1–4 phylogeny and the RSCU data.frame and then created a matrix of phylogenetic proximities between the tips of inferred phylogeny for each DENV1–4 dataset with the function "*proxTips*" and the method oriAbouheif. Finally, the function "*abouheif.moran*" and the method oriAbouheif were applied for the DENV1–4 phylo4d objects and DENV1–4 proximity phylogenetic matrix to calculate Abouheif's Cmean. The package adephylo [35], containing the functions "*proxTips*" and "*abouheif.moran*," and the package phylobase containing the phylo4d constructors are both for R. To perform Blomberg's K test, we employed the function "*multiPhylosignal*" of the package picante for R [36]. This function allows the calculation of phylogenetic dependence for the RSCU DENV1–4 data.frame. We firstly resolved multifurcations (nodes of the tree with two or more descending branches) of the inferred DENV1–4 phylogenetic trees with branches of zero lengths using the function "*multi2di*" of the package ape [27]. In the tests of Abouheif's Cmean and Blomberg's K, the observed codon values for each DENV1–4 dataset were randomly permutated through the tips of each DENV1–4 tree and calculated the focal indices on the new, randomized codon pattern. The repetition of the process for 999 times produced a distribution of the focal indices under random codon usage variation. In comparison of the observed values with these random codon distributions, we took out the quantiles from the tested indices. The quantiles superior than 0.95 for significance level of 0.05 were considered [34].

3. Results

3.1. The G+C Content and ENC. The overall G+C patterns at three nucleotide positions of codons were distinct for each DENV serotype (Figure 1(a), Table S1). The percentage of GC at the first nucleotide position of codon, GC1, is always the highest and that of GC2 is the lowest. GC1 in Asia was the highest in comparison to the other regions. The total GC showed variability among the serotypes (Figure 1(a)) and a characteristic pattern was observed for each serotype. GC3 was more variable than GC1 or GC2 in general and its change was in expense of GC content at preceding positions, particularly GC2. The GC3 value was very close to the total GC and also showed high relationship to it (Table S2). DENV4 had higher GC3 than other serotypes. Meanwhile, the variation profile in GC content among genomes within a DENV serotype was apparently related to their geographic origin (Figure 1(a)).

A matrix correlation analysis with the total GC content and the GC at the three nucleotide positions of codons is shown in Table S2. The total GC content showed a strong correlation with the GC3 ($r \geq 0.7$, $P = 0.01$) for DENV1–4. GC1 had also a strong correlation with GC2 and GC3 in DENV1.

The ENC was also analyzed for each serotype (Table S1). DENV2 appeared with the highest codon bias with a mean 48.8 ± 0.28 whereas DENV4 showed the lowest bias mean (50.87 ± 0.17). ENC bias among genomes within a DENV serotype was correlated with their geographic origin (Figure 1(b), red line). The ENCp analysis showed that the four serotypes had a homogenous codon bias in contrast to ENC (Figure 1(b), blue line). A curve of ENC and ENCp values for each DENV1–4 genome versus their corresponding GC3s data is shown in Figures 2(a) and 2(b). All points of the genome coding sequences lay below the predictable curve. The correlation analysis of ENC and ENCp showed almost no correlation with GC at any of the three codon positions for all DENV1–4 (Table S2). These results indicate that, independent of compositional constraint, some other factors that affect the codon usage variations exist.

3.2. Preferred Codons. The mean and standard deviation of RSCU for 18 amino acids except Met, Trp, and stop codons were determined for each serotype (Table S3). Eleven preferred codons, AGA(Arg), AAC(Asn), GAC(Asp),

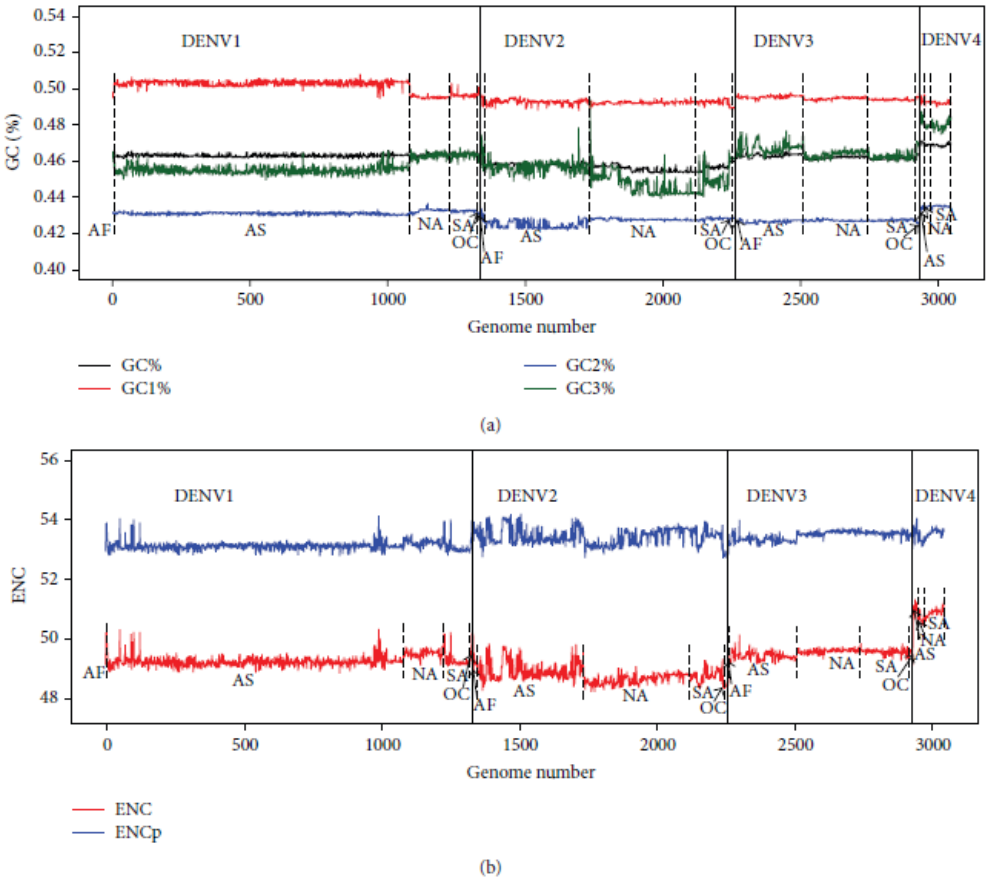


FIGURE 1: The nucleotide composition (G+C) and ENC, ENCp for the 3047 DENV1–4 genomes tested. (a) Total GC and GC content at the three codon positions for each genome. (b) ENC and ENCp for each genome. The dashed lines in both figures indicate the geographical separation within a DENV serotype. The abbreviations mean the following: AF, Africa; AS, Asia; NA, North America; SA, South America; OC, Oceania.

GAA(Glu), GGA(Gly), ATA(Ile), AAA(Lys), CCA(Pro), TCA(Ser), ACA(Thr), and GTG(Val) were consistently shared for all the four DENV (highlighted in blue). There was no extreme bias in preferred codons among specific serotypes. Although in some cases we observed the preferred codons for specific serotypes, these codons still belonged to the set of codons mainly used for the other serotypes. For example, DENV1 used more commonly TAT(Tyr) codon instead of the preferred TAC(Tyr) by DENV2–4. The CAC(His), not CAT(His), codon was preferred in DENV1 and DENV3 while DENV2 and DENV4 use these two codons Tyr and His at proximate frequency. Codon CTG(Leu) was preferred by DENV1 and DENV2, but DENV3 and DENV4 preferred the codon TTG(Leu).

3.3. Correspondence Analysis. One factorial axis accounted for 41.8%, 39.6%, and 40.9% of the total variability in DENV1–3, respectively, indicating that one factor was predominant for those serotypes while for DENV4 dataset the first axis accounted for 25%. The first two axes accounted for more than half of that variability (53–56%) for DENV1–3 except for DENV4 (41%). Thus, the first two factorial axes contribute to the principal differences in codon usage for DENV datasets.

The factor maps produced by crossing axes with the major sources of variation showed well-demarcated geographic separation. They exhibit the following features: (1) in the first axis, as the most important factor on the maps for each serotype (Figures 3(a)–3(d)), the genomes were divided into clusters according to their geographic origin; (2) the Asian, African, and Oceanic genome sequences tend to cluster together; (3) the North American and South American genomes clustered together; (4) the Asian genomes appeared more dispersed than those from other regions; (5) the genomes from other hosts (mosquito, monkey, and unknown host) also clustered accordantly with their geographic sites of isolation. DENV2 showed the most complex geographic pattern. On the other hand, there were also some noticeable "outliers" in the figure, that is, the genomes that were not

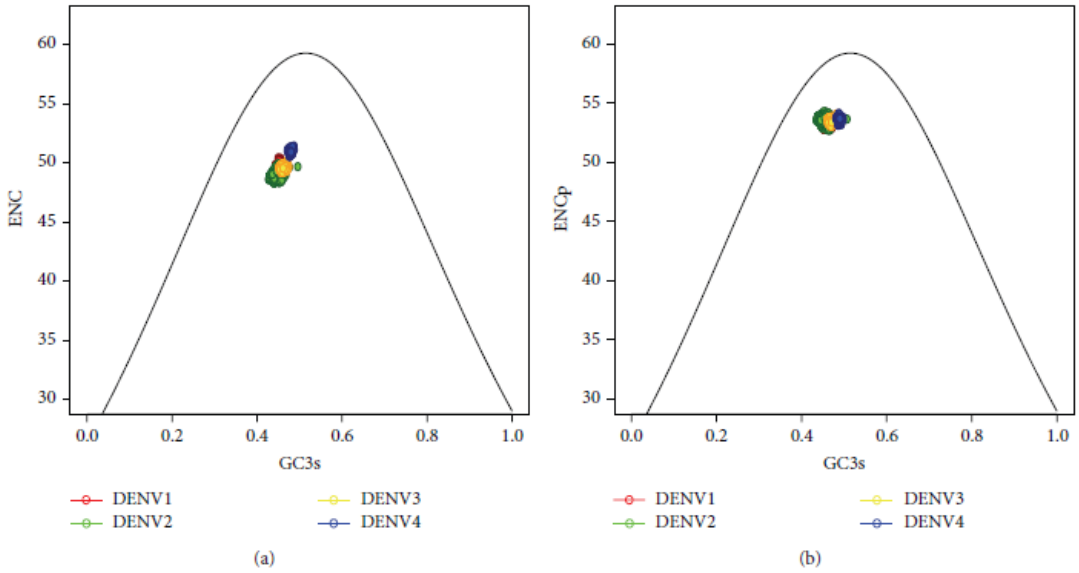


FIGURE 2: Effective number of codons versus GC3s plot of genomes of DENV1–4. (a) ENC versus GC3s, (b) ENCp versus GC3s.

located in the cluster with the majority of strains sharing the same geographic origin (Figure 3, Table S4). Some of these strains have been previously mentioned outside the general cluster in their phylogenetic analysis. For example, the genome from Djibouti, Africa, of serotype 1, previously observed more closely related to the Asian strains due to the existence of a recombinant sequence region with a strain from Singapore (Asia) [37], was located on the Asian cluster in our analysis. Based on this finding we tested if the other identified outliers are recombinant strains with the software RAT. However, no sign of recombination was detected.

The correlations of the GC, GC1, GC2, GC3, ENC, and ENCp of each genome with its position on the first axis are shown in Table 1. Depending on the specific serotype, the genome position on the first axis had strong correlation with GC1 and GC3 for DENV1 and with GC2 and GC3 for DENV2 while others showed less correlation. It is interesting that GC3 showed negative relation with the first axis of the major variation for DENV1 but showed positive correlation with DENV2. ENC and ENCp showed no important correlation with all DENV1–4.

3.4. Phylogenetic and Hierarchical Clustering-Based Trees. The Hierarchical Clustering-Based Trees (HCbT) resulting from the RSCU data are shown in Figures S1(a)–(d). The HCbT revealed two major clusters in each serotype virus. The clusters consisting of Asian, African, and Oceanic genome sequences tend to group together, whereas the clusters enclosing South and North American sequences assemble together. However, some Asian genomes were located at the clusters of North and South American strains. The genomes identified as outliers were also located in the same geographical clusters as indicated in our CA analysis. On the other hand, the phylogenetic relationships among DENV genomes were also constructed based on the genome nucleotide sequences (Figures S1(e)–(h)). The comparison of these phylogenetic trees showed that these analyses on two datasets showed similar results. Moreover, the majority of the outliers were also confirmed by the inferred phylogenetic trees.

3.5. Evaluation of Phylogenetic Dependence of Codon Usage. The 59 codons usage values for individual genome in each DENV dataset were tested for phylogenetic dependence. We followed Abouheif's Cmean approach. The null hypothesis of lacking phylogenetic dependence was rejected ($P = 0.05$) for all 59 codon variables with Abouheif's Cmean statistic for DENV1–3 (Table S5). In DENV4, the absence of phylogenetic autocorrelation was not significantly rejected for the following codons: CGT, CTG, TAC, TAT, TTC, TTG, and TTT. However, although the phylogenetic dependence varied across the 59 codons, the null hypothesis of lacking phylogenetic dependence was rejected for all DENV1–4 by means of Blomberg's K statistic (Table S6). The statistical results indicate the presence of phylogenetic dependence of codon usage in DENV genomes.

4. Discussion

The identification of principal factors shaping codon usage is important for understanding the evolution of organisms, including viruses. In the present study, the analysis of total GC relation with the three nucleotide positions of codons GC1, GC2, and GC3 showed that the forces shaping codon usage were not the same for all codon positions (Table S2). The GC3 had the highest correlation with total GC and was very close to the total GC value in DENV1–4, suggesting a

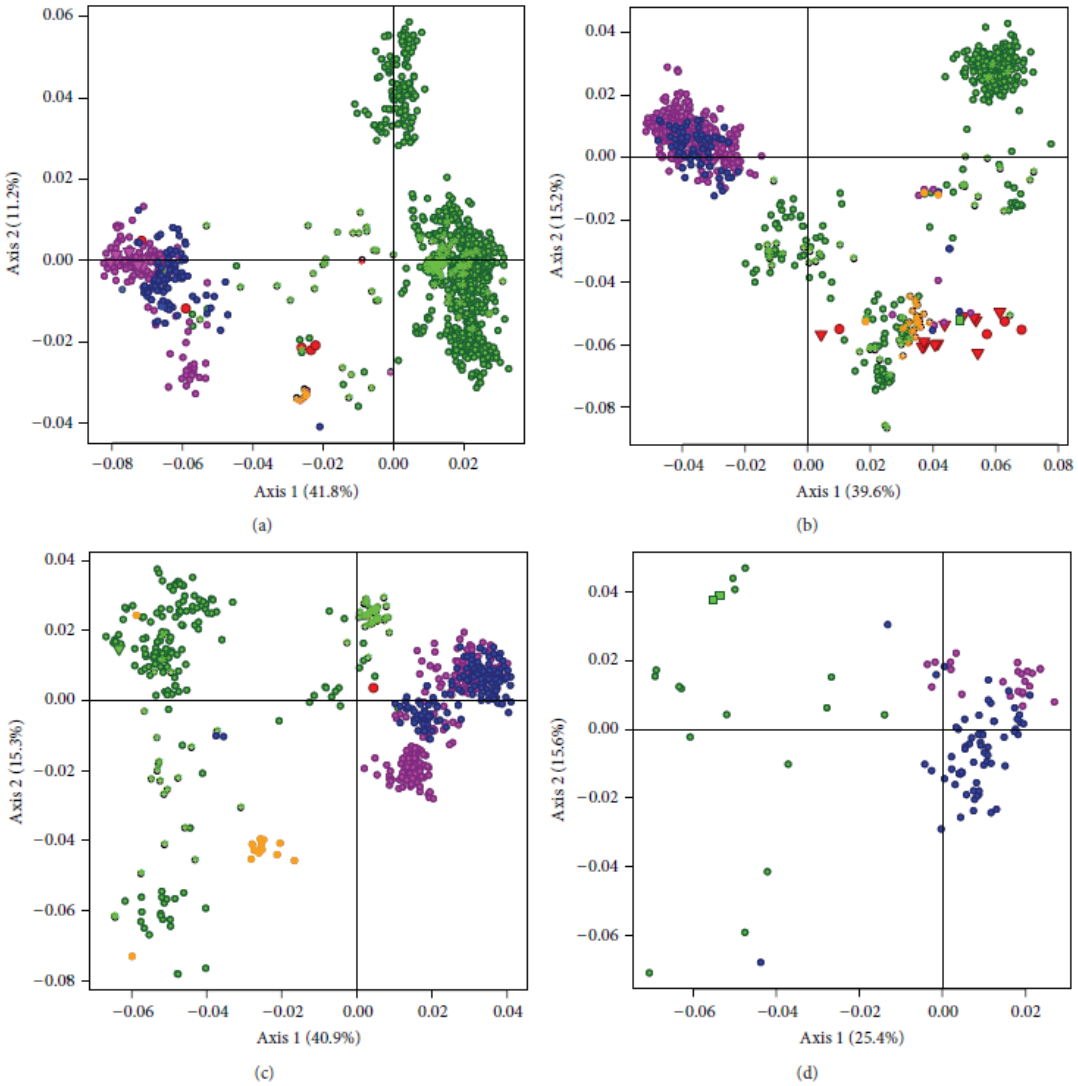


FIGURE 3: Correspondence analysis based on RSCU values for DENV. The geographic regions of isolates are indicated in colors as red (African), green (Asian), magenta (North American), blue (South American), and orange (Oceanic). The host sources are, respectively, represented as circles for human, squares for monkey, inverted triangles for mosquito, and asterisks for unknown host. (a) DENV1; (b) DENV2; (c) DENV3; (d) DENV4.

TABLE 1: The correlation analysis of GC, ENC, and ENCp with the first axis of major variation.

Serotype	A1* % of variation	GC(r)	GC1(r)	GC2(r)	GC3(r)	ENC(r)	ENCp(r)
DENV1	41.8	−0.40	0.87	−0.54	−0.88	−0.47	−0.18
DENV2	39.8	0.57	0.00	−0.79	0.78	0.19	−0.18
DENV3	40.9	−0.55	−0.55	0.61	−0.59	0.44	0.59
DENV4	25.4	−0.40	−0.37	0.66	−0.46	−0.55	−0.50

* represents axis 1 in the correspondence analysis.

strong mutational pressure on the third position of codons. The ENC or ENCp versus GC3 plots showed that, in addition to compositional constraint, some other factors have effect on the codon usage variations. GC2 does not have important correlation with total GC in the examined genomes in the present study, implying that the constraint on this codon position is possibly due to the functional selection. A recent paper showed that the mutations on this position in the analyzed samples were mostly nonsynonymous substitutions [15]. These results demonstrated that both mutational and purifying selection pressures are the major forces in influencing the codon usage among DENV, consistent with some previous reports [6, 38], but these factors have distinct pressure on specific nucleotide position of a codon.

The analysis of ENC showed an overall weak codon usage bias, as shown in Table S1, where DENV2 has the highest codon bias (48.80) and DENV4 has the lowest one (50.87). This result is similar to a recent report [14], indicating that the result was not affected by an increased number of samples and might represent an inherent feature of DENV. One plausible explanation could be that DENV4 is less adapted to human environment, whereas DENV2 is more adapted to humans. On the other hand, DENV2 has been associated with more aggressive diseases forms and is generally the most prevailing serotype during outbreaks situations [39]. These could mean that codon bias of DENV2 contributes to successful infection in human cells in comparison with DENV4.

Moreover, the CA and HCbT analyses within each serotype showed similar clustering patterns for the four serotypes. The DENV strains occurring in the same continental region are more closely related, forming a cluster, indicating that viruses from a geographical group show similar codon usage bias. The Asian genomes of the four serotypes showed a wide diversity in the clusters and each of them can be further divided into more homogenous subgroups. This more diversified clustering could be the consequence of longer times of DENV evolution in Asia than in other regions. Some of the Asian genomes clustered close to the American ones, implying an evolutionary link between the Asian and American clusters. The North and South American strains tend to cluster more homogenously together with less codon usage variations, corresponding to the previous observation that a limited nucleotide diversity exists in American DENV strains [1, 15]. As the DENV in North and South America came from Asia, the homogenous cluster in North and South American populations could indicate a simple event of introduction from Asia, then spreading over this continent with much less adaptation time than in Asia, as the consequence of founder effect.

The sequences isolated from mosquito and monkey genomes in the CA were also grouped with human strains from the same geographic origin, indicating that sylvatic DENV changes in adaptation on codon usage in a similar way to endemic human DENV, as indicated by the study on nucleotide sequences [40]. On the other hand, Zhou et al. reported no link of geographic origin to the codon usage of DENV [13]. Behura and Severson found that the silent sites are favoring the geographical diversification [15]. Our study showed that not only GC3 but also GC1 and GC2 have a good correlation with axis major variation, depending on the serotype, suggesting that all the codon sites are related to clustering of geographical strains. Thus, the present study demonstrated the strong influence of geographic origin of DENV on shaping codon usage patterns. The discrepancy in results from studies may be due to the magnitude of samples used for analysis.

The clustering groups based on the codon usage datasets or phylogenetic tree on nucleotide sequence dataset showed the similar clustering results. This observation indicates the influence of the species evolution of DENV at the level of codon usage. We applied two statistical methods to assess the phylogenetic dependence of codon usage values. The positive results suggested that codon usage of DENV is engaged in the evolution of DENV lineages. The phylogenetic dependence is often interpreted as an information provider on the evolutionary process or rate [32]. For instance, it is common to associate the lack of phylogenetic dependence with evolutionary lability and the presence of phylogenetic dependence with stabilizing selection. Thus, the phylogenetic dependence analysis in the present study suggests that stabilizing selection acts on codon bias.

In summary, the codon usage of DENV genomes was analyzed on a large scale. Our analysis demonstrated that both mutational and purifying selection pressures have important contribution to the codon usage; however, these factors have distinct pressure on specific codon nucleotide positions. The codon usage patterns of DENV genomes showed apparent geographic feature. The phylogenetic dependence analysis suggests that stabilizing selection acts on codon bias.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Consejo Nacional de Ciencia y Tecnología, México (http://www.conacyt.mx/) (Fondo Sectorial de Investigación Básica SEP CONACyT (CB-2011-01) with Grant no. 168541) and Secretaría de Investigación y Posgrado del Instituto Politécnico Nacional, México (http://www.sip.ipn.mx/WPS/WCM/CONNECT/SIP/SIP/INICIO/INDEX.HTM) (Grant no. SIP20130400). Xianwu Guo, Ma Isabel Salazar, and Juan Salas Benito hold a scholarship from Comisión de Operación y Fomento de Actividades Académicas/Instituto Politécnico Nacional.

References

[1] P. Rivera-Osorio, G. Vaughan, J. E. Ramírez-González et al., "Molecular epidemiology of autochthonous dengue virus strains circulating in Mexico," *Journal of Clinical Microbiology*, vol. 49, no. 9, pp. 3370–3374, 2011.

[2] D. Normile, "Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts," *Science*, vol. 342, no. 6157, p. 415, 2013.

[3] N. Vasilakis, J. Cardosa, K. A. Hanley, E. C. Holmes, and S. C. Weaver, "Fever from the forest: prospects for the continued emergence of sylvatic dengue virus and its impact on public health," *Nature Reviews Microbiology*, vol. 9, no. 7, pp. 532–541, 2011.

[4] WHO, "Impact of dengue," 2014, http://www.who.int/csr/disease/dengue/impact/en/.

[5] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé, "Codon catalog usage and the genome hypothesis," *Nucleic Acids Research*, vol. 8, no. 1, pp. r49–r62, 1980.

[6] G. M. Jenkins and E. C. Holmes, "The extent of codon usage bias in human RNA viruses and its evolutionary origin," *Virus Research*, vol. 92, no. 1, pp. 1–7, 2003.

[7] F. P. Lobo, B. E. F. Mota, S. D. J. Pena et al., "Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts," *PLoS ONE*, vol. 4, no. 7, Article ID e6282, 2009.

[8] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 725–736, 1994.

[9] G. M. Jenkins, M. Pagel, E. A. Gould, P. M. de A Zanotto, and E. C. Holmes, "Evolution of base composition and codon usage bias in the genus *Flavivirus*," *Journal of Molecular Evolution*, vol. 52, no. 4, pp. 383–390, 2001.

[10] A. M. Schubert and C. Putonti, "Evolution of the sequence composition of *Flaviviruses*," *Infection, Genetics and Evolution*, vol. 10, no. 1, pp. 129–136, 2010.

[11] B. K. Rima and N. V. McFerran, "Dinucleotide and stop codon frequencies in single-stranded RNA viruses," *Journal of General Virology*, vol. 78, no. 11, pp. 2859–2870, 1997.

[12] M.-W. Su, W. C. Chu, and H. S. Yuan, "Distinguish dengue virus serotypes via codon usage patterns," in *Proceedings of the 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE '07)*, pp. 1328–1330, Wuhan, China, July 2007.

[13] J. H. Zhou, J. Zhang, D. J. Sun et al., "The distribution of synonymous codon choice in the translation initiation region of dengue virus," *PLoS ONE*, vol. 8, no. 10, Article ID e77239, 2013.

[14] J. J. Ma, F. Zhao, J. Zhang et al., "Analysis of synonymous codon usage in dengue viruses," *Journal of Animal and Veterinary Advances*, vol. 12, no. 1, pp. 88–98, 2013.

[15] S. K. Behura and D. W. Severson, "Nucleotide substitutions in dengue virus serotypes from Asian and American countries: insights into intracodon recombination and purifying selection," *BMC Microbiology*, vol. 13, article 37, no. 1, 2013.

[16] W. Resch, L. Zaslavsky, B. Kiryutin, M. Rozanov, Y. Bao, and T. A. Tatusova, "Virus variation resources at the National Center for Biotechnology Information: dengue virus," *BMC Microbiology*, vol. 9, article 65, 2009.

[17] P. M. Sharp and W. Li, "The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[18] F. Wright, "The "effective number of codons" used in a gene," *Gene*, vol. 87, no. 1, pp. 23–29, 1990.

[19] J. A. Novembre, "Accounting for background nucleotide composition when measuring codon usage bias," *Molecular Biology and Evolution*, vol. 19, no. 8, pp. 1390–1394, 2002.

[20] D. Charif and J. R. Lobry, "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis structural approaches to sequence evolution," in *Structural Approaches to Sequence Evolution*, U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Eds., pp. 207–232, Springer, Berlin, Germany, 2007.

[21] R-Development-Core-Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[22] J. F. Peden, *Analysis of Codon Usage*, 1999, http://codonw.sourceforge.net/.

[23] M. Greenacre, *Correspondence Analysis in Practice*, Chapman & Hall/CRC, London, UK, 2nd edition, 2007.

[24] S. Dray and A. B. Dufour, "The ade4 package: implementing the duality diagram for ecologists," *Journal of Statistical Software*, vol. 22, no. 4, pp. 1–20, 2007.

[25] R. Taylor, "Interpretation of the correlation coefficient: a basic review," *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, 1990.

[26] H. Suzuki, C. J. Brown, L. J. Forney, and E. M. Top, "Comparison of correspondence analysis methods for synonymous codon usage in bacteria," *DNA Research*, vol. 15, no. 6, pp. 357–365, 2008.

[27] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.

[28] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[29] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2—approximately maximum-likelihood trees for large alignments," *PLoS ONE*, vol. 5, no. 3, Article ID e9490, 2010.

[30] M. Worobey and E. C. Holmes, "Evolutionary aspects of recombination in RNA viruses," *Journal of General Virology*, vol. 80, no. 10, pp. 2535–2543, 1999.

[31] G. J. Etherington, J. Dicks, and I. N. Roberts, "Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination," *Bioinformatics*, vol. 21, no. 3, pp. 278–281, 2005.

[32] L. J. Revell, L. J. Harmon, and D. C. Collar, "Phylogenetic signal, evolutionary process, and rate," *Systematic Biology*, vol. 57, no. 4, pp. 591–601, 2008.

[33] S. K. Behura, B. K. Singh, and D. W. Severson, "Antagonistic relationships between intron content and codon usage bias of genes in three mosquito species: functional and evolutionary implications," *Evolutionary Applications*, vol. 6, no. 7, pp. 1079–1089, 2013.

[34] T. Münkemüller, S. Lavergne, B. Bzeznik et al., "How to measure and test phylogenetic signal," *Methods in Ecology and Evolution*, vol. 3, no. 4, pp. 743–756, 2012.

[35] T. Jombart, F. Balloux, and S. Dray, "adephylo: new tools for investigating the phylogenetic signal in biological traits," *Bioinformatics*, vol. 26, no. 15, pp. 1907–1909, 2010.

[36] S. W. Kembel, P. D. Cowan, M. R. Helmus et al., "Picante: R tools for integrating phylogenies and ecology," *Bioinformatics*, vol. 26, no. 11, pp. 1463–1464, 2010.

[37] H. J. G. Tolou, P. Couissinier-Paris, J.-P. Durand et al., "Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome suquences," *Journal of General Virology*, vol. 82, no. 6, pp. 1283–1290, 2001.

[38] E. C. Holmes, "Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus," *Journal of Virology*, vol. 77, no. 20, pp. 11296–11298, 2003.

[39] R. Cologna, P. M. Armstrong, and R. Rico-Hesse, "Selection for virulent dengue viruses occurs in humans and mosquitoes," *Journal of Virology*, vol. 79, no. 2, pp. 853–859, 2005.

[40] N. Vasilakis, E. C. Holmes, E. B. Fokam et al., "Evolutionary processes among sylvatic dengue type 2 viruses," *Journal of Virology*, vol. 81, no. 17, pp. 9591–9595, 2007.