DNA AND CELL BIOLOGY
Volume 36, Number 2, 2017
© Mary Ann Liebert, Inc.
Pp. 1–8
DOI: 10.1089/dna.2016.3532

ORIGINAL RESEARCH ARTICLE

Zika and Flaviviruses Phylogeny Based on the Alignment-Free Natural Vector Method

Yongkun Li,[1] Lily He,[1] Rong Lucy He,[2] and Stephen S.-T. Yau[1]

Zika virus (ZIKV) is a mosquito-borne flavivirus. It was first isolated from Uganda in 1947 and has become an emergent event since 2007. However, because of the inconsistency of alignment methods, the evolution of ZIKV remains poorly understood. In this study, we first use the complete protein and an alignment-free method to build a phylogenetic tree of 87 Zika strains in which Asian, East African, and West African lineages are characterized. We also use the NS5 protein to construct the genetic relationship among 44 Zika strains. For the first time, these strains are divided into two clades: African 1 and African 2. This result suggests that ZIKV originates from Africa, then spread to Asia, Pacific islands, and throughout the Americas. We also perform the phylogeny analysis for 53 viruses in genus *Flavivirus* to which ZIKV belongs using complete proteins. Our conclusion is consistent with the classification by the hosts and transmission vectors.

Keywords: Zika virus, *Flavivirus*, evolution, natural vector

Introduction

Zika virus (ZIKV) is an arbovirus within the genus *Flavivirus* and family Flaviviridae (Baronti *et al.*, 2014; Enfissi *et al.*, 2016). It was transmitted by *Aedes* (Stegomyia) mosquitoes and closely related to other flaviviruses such as dengue, West Nile, and yellow fever viruses. ZIKV was first identified in 1947 in Uganda and since then, it caused sporadic human infections throughout Africa and Asia. The first large epidemic was reported on Yap Island, Micronesia in 2007 (Lanciotti *et al.*, 2008; Duffy *et al.*, 2009). Then it had an outbreak in French Polynesia in 2013–2014, during which a dengue epidemic and increase in severe neurological complications were reported (Cao-Lormeau *et al.*, 2014).

Until 2015, the strain of ZIKV emerged in the city of Natal in Brazil (Zanluca *et al.*, 2015). Soon after, autochthonous transmission of ZIKV took place throughout other countries in the Americas. After the outbreak in Brazil, a significant increase of microcephaly in newborns was documented by the Brazilian Ministry of Health. Some authors propose that the microcephaly is possibly caused by ZIKV (Broutet *et al.*, 2016). Humans infected with ZIKV present a dengue-like syndrome related to mild fever, headache, cutaneous rashes, retroorbital pain, and conjunctivitis (Faye *et al.*, 2014). This clinical characteristic could easily be mistaken for dengue or chikungunya fevers, which are two common diseases associated with public health.

ZIKV is a positive-sense single-stranded RNA virus, with an $\sim 11\,\text{kb}$ genome. The genome contains a single open reading frame (ORF) with flanking 5′ and 3′ untranslated regions. The ORF encodes a polyprotein that is cleaved into three structural proteins: the capsid, precursor of membrane, and envelope (E), and seven nonstructural proteins: NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5. Among them, the NS5 protein is the largest viral protein, of which C-terminus is related to RNA polymerase activity and the N-terminus is involved in RNA capping.

As far as we know, the phylogenetic trees of ZIKV were all constructed by multiple sequence alignment algorithms and based on genomic sequences. However, these trees were inconsistent with one another. For example, the recovered evolutionary relationship presented two cases: Asian and African lineages; Asian and two African lineages. Based on the complete genomes, the two different cases of ZIKV phylogeny could be obtained (Enfissi *et al.*, 2016; Faria *et al.*, 2016; Lanciotti *et al.*, 2016). Using the E and NS5 genes, although some studies generated three lineages, the topology of trees was different (Faye *et al.*, 2014; Shen *et al.*, 2016). The phylogeny derived from partial genes may be two lineages as well (Zhu *et al.*, 2016). Moreover, the proposed hypothesis that ZIKV is originated from Africa has not been verified (Gong *et al.*, 2016). The protein sequences of ZIKV are directly involved in a variety of biological processes and more conserved than the gene sequences. To fully understand the origin and diversity of this emergent agent, it is necessary to perform phylogeny analysis by alignment-free methods with protein sequences.

[1]Department of Mathematical Sciences, Tsinghua University, Beijing, People's Republic of China.
[2]Department of Biological Sciences, Chicago State University, Chicago, Illinois.

Over the past decade, the number of alignment-free methods for inferring phylogeny of organisms increased sharply. Many of them are variants of k-word methods, which only take advantage of the occurrence of words of length k (Dai *et al.*, 2008; Wu *et al.*, 2009). The recently proposed natural vector method includes information from both positions and counts of amino acids (Deng *et al.*, 2011). As shown in the previous work (Yu *et al.*, 2013), the natural vector method can achieve more accurate results than alignment-based methods for phylogeny analysis. The method is also fast in computational speed compared with alignment-based ones.

The investigation for other flaviviruses in the genus *Flavivirus*, to which ZIKV belongs, has been limited due to the lack of sequenced genomes and their inability to infect vertebrates. In recent years yet, some of the viruses caused several global epidemics or became localized public health concern. These viruses are small, enveloped, single-stranded and positive-sense ssRNA viruses, which infect various hosts such as birds, mammals and insects. Their genomes are ~ 11 kb in length and have a similar structure to ZIKV (Billoir *et al.*, 2000; Moureau *et al.*, 2015). Due to the wide diversity among these flaviviruses, knowledge of their genetic relationships is limited and mainly comes from alignment-based methods. In this study, the evolutionary relationships of ZIKV strains and other flaviviruses are constructed by the natural vector methods with protein sequences. We also verify the hypothesis for the Africa origin of ZIKV.

Materials and Methods

Source of datasets

There is only one reference genome from ZIKV in NCBI and many ZIKV strains have short partial genes. Some complete genomes of ZIKV strains also show difference in length. Since the complete coding region of ZIKV encodes a long protein (polyprotein), which is subsequently processed into 10 proteins, including the major NS5 protein, the polyprotein sequences of 87 ZIKV strains were retrieved from NCBI on 23 June 2016. The accession numbers, countries where the viruses are isolated, the collection dates, and length of polyproteins are listed in Supplementary Table S1; Supplementary Data are available online at www.liebertpub.com/dna.

The proteins translated from the NS5 gene were obtained from the 87 complete proteins. Among them 43 protein sequences are identical to others and then only the remaining 44 protein sequences are kept in subsequent analysis. The polyprotein sequences of 53 referenced genomes in the *Flavivirus* genus were extracted from NCBI as well, and their accession numbers are listed in Supplementary Table S2.

Natural vector

The recently proposed natural vector method is a powerful tool for virus classification and phylogeny (Deng *et al.*, 2011). This method is alignment free and needs no parameters. By this approach, each protein sequence is mapped into a 60-dimensional numeric vector, which has been proven to perform well to capture information in viral sequences (Yu *et al.*, 2013; Huang *et al.*, 2014; Li *et al.*, 2016). We review the method as follows. Let Ω be the set of 20 types of amino acids, that is, $\Omega = \{A, R, N, D, \ldots, Y, V\}$, $S = (s_1, s_2, \ldots, s_n)$ be a protein sequence of length $_n$. For $k \in \Omega$, define $w_k(\cdot) : \Omega \rightarrow \{0, 1\}$ such that $w_k(s_i) = 1$ if $s_i = k$ and 0 otherwise.

(1) Let $n_k = \sum\limits_{i=1}^{n} w_k(s_i)$ denote the number of letter k in S.

(2) Let $\mu_k = \sum\limits_{i=1}^{n} \frac{i}{n_k} \cdot w_k(s_i)$ be the average position of letter k.

(3) Let $D_2^k = \sum\limits_{i=1}^{n} \frac{(i-\mu_k)^2}{n_k \cdot n} \cdot w_k(s_i)$ be the scaled second central moment of positions of letter k.

For ambiguous amino acids, the letter J indicates it may be L or I; B for D or N; Z for Q or E; and X for all possible 20 types of amino acids. Thus, for $k \in \Omega$ we define the weight $w_k(s_i)$ as the expected number of occurrence of letter k in position i. For example, $w_L(s_i) = 1$ if $s_i = L$; $w_L(s_i) = 0.5$ if $s_i = J$; $w_L(s_i) = 0.05$ if $s_i = X$; $w_L(s_i) = 0$ for other s_i.

Thus, the 60-dimensional natural vector of a protein sequence S is defined as

$$(n_A, \ldots, n_V, \mu_A, \ldots, \mu_V, D_2^A, \ldots, D_2^V).$$

The Euclidean distance is a true mathematical distance that satisfies the triangle inequality, that is, $d(A, B) \leq d(A, C) + d(C, B)$. Here, we define $d(A,B)$ as

$$d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots (a_m - b_m)^2},$$

where $A = (a_1, a_2, \ldots, a_m), B = (b_1, b_2, \ldots, b_m)$, m is the dimension of A and B vectors. In our study, A and B are two 60-dimensional natural vectors and m is equal to 60. Once the natural vectors of viral protein sequences are acquired, pairwise distance between all viruses is generated with the Euclidean distance. A distance matrix is built and then the phylogeny of these viruses can be reconstructed with the neighbor-joining algorithm packaged in MEGA 6.0 software.

Results

Phylogenetic analysis of Zika strains

According to the organization of ZIKV genome, the complete coding region of ZIKV is first translated into a protein (polyprotein) which is then cleaved into three structural and seven nonstructural proteins. Based on the 60-dimensional natural vector method, the phylogenetic tree using 87 ZIKV polyproteins is reconstructed. As illustrated in Figure 1, the 87 ZIKV are well classified into three lineages: the Asian, West African, and East African lineages, which is consistent with the previous work (Lanciotti *et al.*, 2016). The two African clades form a sister group to the Asian clade. The newly identified ZIKV strains in countries of Americas (such as Brazil, Haiti, Suriname, Guatemala, Martinique, and Puerto Rico) are all close to Asian and Pacific strains as shown in Figure 1.

The strains from the Asian and Pacific countries are in the base of the Asian clade consisting of strains from Asia, Pacific regions, and Americas. This indicates that the

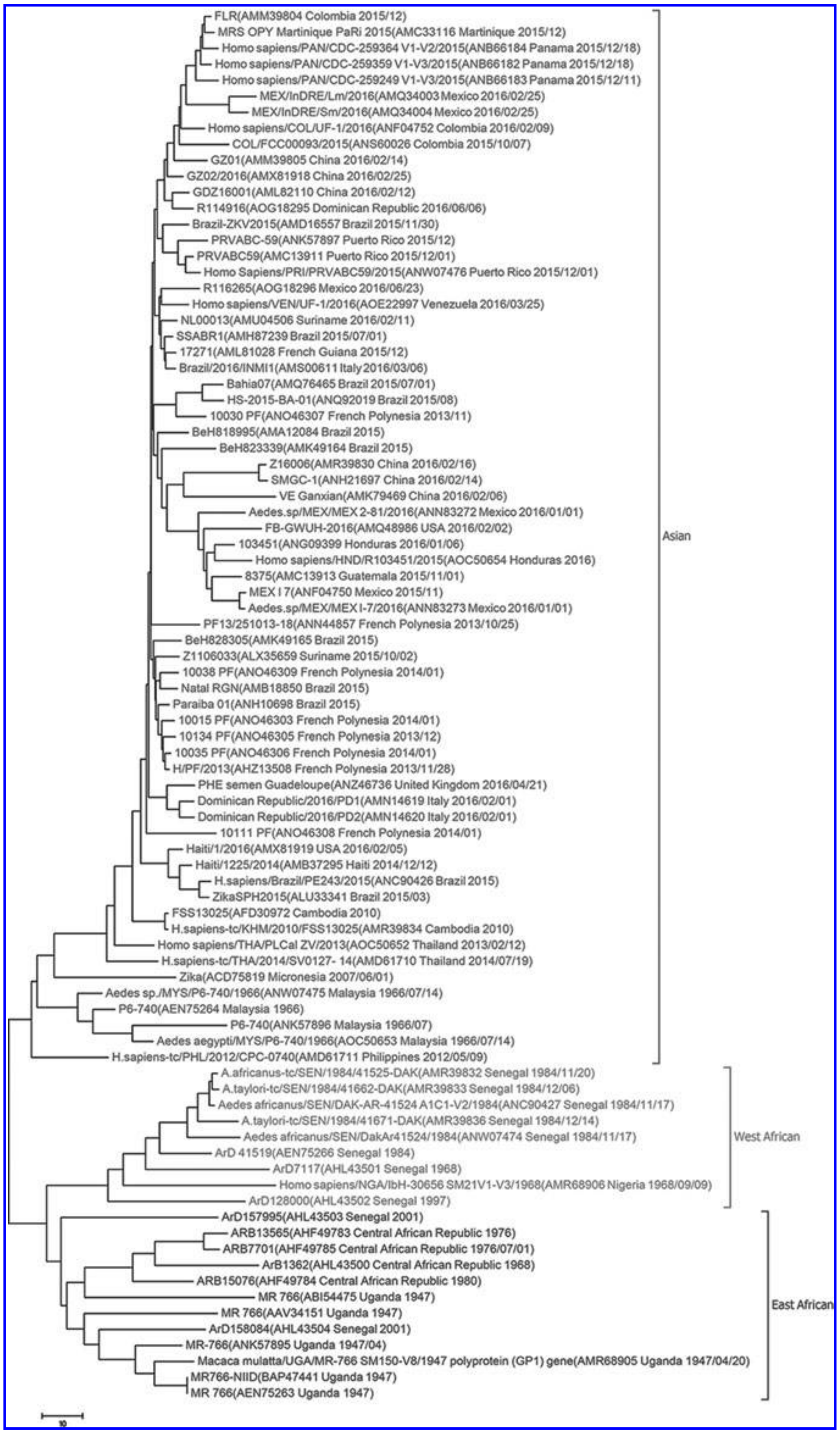
FLR(AMM39804 Colombia 2015/12)
MRS OPY Martinique PaRi 2015(AMC33116 Martinique 2015/12)
Homo sapiens/PAN/CDC-259364 V1-V2/2015(ANB66184 Panama 2015/12/18)
Homo sapiens/PAN/CDC-259359 V1-V3/2015(ANB66182 Panama 2015/12/18)
Homo sapiens/PAN/CDC-259249 V1-V3/2015(ANB66183 Panama 2015/12/11)
MEX/InDRE/Lm/2016(AMQ34003 Mexico 2016/02/25)
MEX/InDRE/Sm/2016(AMQ34004 Mexico 2016/02/25)
Homo sapiens/COL/UF-1/2016(ANF04752 Colombia 2016/02/09)
COL/FCC00093/2015(ANS60026 Colombia 2015/10/07)
GZ01(AMM39805 China 2016/02/14)
GZ02/2016(AMX81918 China 2016/02/25)
GDZ16001(AML82110 China 2016/02/12)
R114916(AOG18295 Dominican Republic 2016/06/06)
Brazil-ZKV2015(AMD16557 Brazil 2015/11/30)
PRVABC-59(ANK57897 Puerto Rico 2015/12)
PRVABC59(AMC13911 Puerto Rico 2015/12/01)
Homo Sapiens/PRI/PRVABC59/2015(ANW07476 Puerto Rico 2015/12/01)
R116265(AOG18296 Mexico 2016/06/23)
Homo sapiens/VEN/UF-1/2016(AOE22997 Venezuela 2016/03/25)
NL00013(AMU04506 Suriname 2016/02/11)
SSABR1(AMH87239 Brazil 2015/07/01)
17271(AML81028 French Guiana 2015/12)
Brazil/2016/INMI1(AMS00611 Italy 2016/03/06)
Bahia07(AMQ76465 Brazil 2015/07/01)
HS-2015-BA-01(ANQ92019 Brazil 2015/08)
10030 PF(ANO46307 French Polynesia 2013/11)
BeH818995(AMA12084 Brazil 2015)
BeH823339(AMK49164 Brazil 2015)
Z16006(AMR39830 China 2016/02/16)
SMGC-1(ANH21697 China 2016/02/14)
VE Ganxian(AMK79469 China 2016/02/06)
Aedes.sp/MEX/MEX 2-81/2016(ANN83272 Mexico 2016/01/01)
FB-GWUH-2016(AMQ48986 USA 2016/02/02)
103451(ANG09399 Honduras 2016/01/06)
Homo sapiens/HND/R103451/2015(AOC50654 Honduras 2016)
8375(AMC13913 Guatemala 2015/11/01)
MEX I 7(ANF04750 Mexico 2015/11)
Aedes.sp/MEX/MEX I-7/2016(ANN83273 Mexico 2016/01/01)
PF13/251013-18(ANN44857 French Polynesia 2013/10/25)
BeH828305(AMK49165 Brazil 2015)
Z1106033(ALX35659 Suriname 2015/10/02)
10038 PF(ANO46309 French Polynesia 2014/01)
Natal RGN(AMB18850 Brazil 2015)
Paraiba 01(ANH10698 Brazil 2015)
10015 PF(ANO46303 French Polynesia 2014/01)
10134 PF(ANO46305 French Polynesia 2013/12)
10035 PF(ANO46306 French Polynesia 2014/01)
H/PF/2013(AHZ13508 French Polynesia 2013/11/28)
PHE semen Guadeloupe(ANZ46736 United Kingdom 2016/04/21)
Dominican Republic/2016/PD1(AMN14619 Italy 2016/02/01)
Dominican Republic/2016/PD2(AMN14620 Italy 2016/02/01)
10111 PF(ANO46308 French Polynesia 2014/01)
Haiti/1/2016(AMX81919 USA 2016/02/05)
Haiti/1225/2014(AMB37295 Haiti 2014/12/12)
H.sapiens/Brazil/PE243/2015(ANC90426 Brazil 2015)
ZikaSPH2015(ALU33341 Brazil 2015/03)
FSS13025(AFD30972 Cambodia 2010)
H.sapiens-tc/KHM/2010/FSS13025(AMR39834 Cambodia 2010)
Homo sapiens/THA/PLCal ZV/2013(AOC50652 Thailand 2013/02/12)
H.sapiens-tc/THA/2014/SV0127- 14(AMD61710 Thailand 2014/07/19)
Zika(ACD75819 Micronesia 2007/06/01)
Aedes sp./MYS/P6-740/1966(ANW07475 Malaysia 1966/07/14)
P6-740(AEN75264 Malaysia 1966)
P6-740(ANK57896 Malaysia 1966/07)
Aedes aegypti/MYS/P6-740/1966(AOC50653 Malaysia 1966/07/14)
H.sapiens-tc/PHL/2012/CPC-0740(AMD61711 Philippines 2012/05/09)
Asian
A.africanus-tc/SEN/1984/41525-DAK(AMR39832 Senegal 1984/11/20)
A.taylori-tc/SEN/1984/41662-DAK(AMR39833 Senegal 1984/12/06)
Aedes africanus/SEN/DAK-AR-41524 A1C1-V2/1984(ANC90427 Senegal 1984/11/17)
A.taylori-tc/SEN/1984/41671-DAK(AMR39836 Senegal 1984/12/14)
Aedes africanus/SEN/DakAr41524/1984(ANW07474 Senegal 1984/11/17)
ArD 41519(AEN75266 Senegal 1984)
ArD7117(AHL43501 Senegal 1968)
Homo sapiens/NGA/IbH-30656 SM21V1-V3/1968(AMR68906 Nigeria 1968/09/09)
ArD128000(AHL43502 Senegal 1997)
West African
ArD157995(AHL43503 Senegal 2001)
ARB13565(AHF49783 Central African Republic 1976)
ARB7701(AHF49785 Central African Republic 1976/07/01)
ArB1362(AHL43500 Central African Republic 1968)
ARB15076(AHF49784 Central African Republic 1980)
MR 766(ABI54475 Uganda 1947)
MR 766(AAV34151 Uganda 1947)
ArD158084(AHL43504 Senegal 2001)
MR-766(ANK57895 Uganda 1947/04)
Macaca mulatta/UGA/MR-766 SM150-V8/1947 polyprotein (GP1) gene(AMR68905 Uganda 1947/04/20)
MR766-NIID(BAP47441 Uganda 1947)
MR 766(AEN75263 Uganda 1947)
East African
10

FIG. 1. Phylogenetic tree of 87 Zika viruses with polyproteins based on the 60-dimensional natural vector method. The strains are divided into three lineages: the Asian lineage, the West African, and East African lineages. Names, accession number, location, and year identified of the strains are displayed.

epidemics in the Americas are likely dated back to strains from Malaysia isolated in 1966. The two African lineages are sister groups. The strains from Senegal are distributed in the two lineages, which suggest that two independent lineages have been circulating in this country. The MR766 prototype strains are clustered together as well.

Using 60-dimensional natural vectors, protein sequences coded by NS5 genes are employed to infer the phylogeny. NS5 is the largest protein in Zika and has the largest number of amino acid substitutions (Zhu *et al.*, 2016). Thus, this protein provides more clues to evolutionary relationship between Zika strains. As displayed in Figure 2, the strains are grouped into two clades: the African 1 and African 2 clades. The recently emergent ZIKV strains in Americas are also near to those from Asia and Pacific islands. This result suggests that the Asian genotype expands to Americas. The tree also indicates that the Malaysia strains identified in 1966 are representative of an ancestral genotype for the Asian clade.

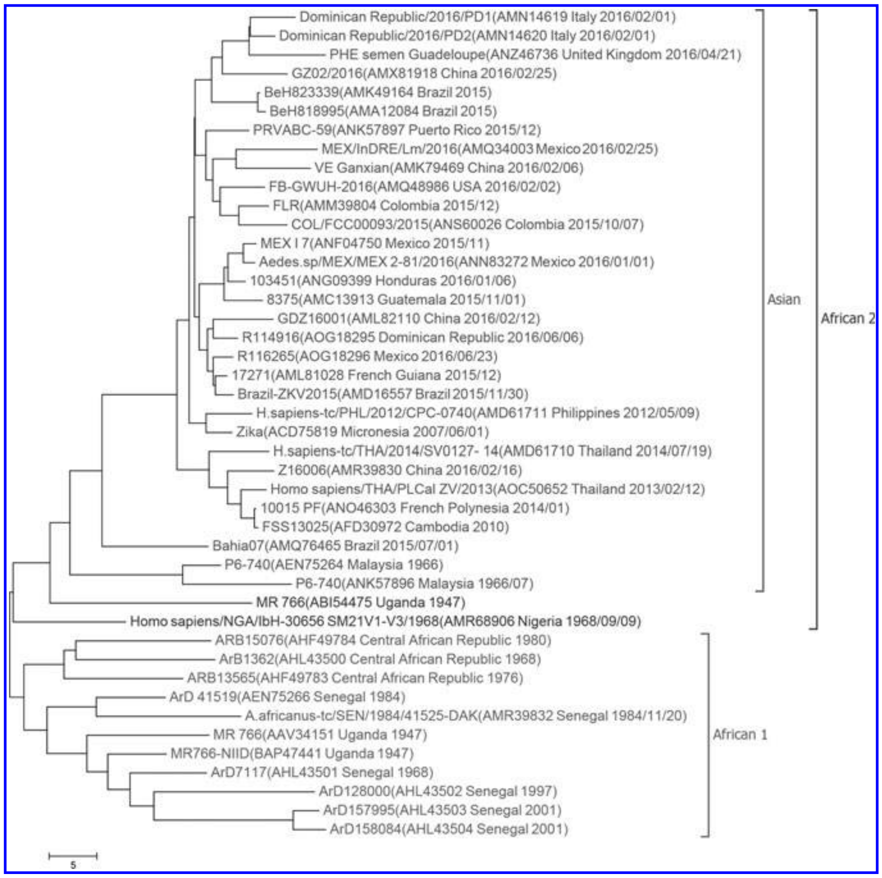
Dominican Republic/2016/PD1(AMN14619 Italy 2016/02/01)
Dominican Republic/2016/PD2(AMN14620 Italy 2016/02/01)
PHE semen Guadeloupe(ANZ46736 United Kingdom 2016/04/21)
GZ02/2016(AMX81918 China 2016/02/25)
BeH823339(AMK49164 Brazil 2015)
BeH818995(AMA12084 Brazil 2015)
PRVABC-59(ANK57897 Puerto Rico 2015/12)
MEX/InDRE/Lm/2016(AMQ34003 Mexico 2016/02/25)
VE Ganxian(AMK79469 China 2016/02/06)
FB-GWUH-2016(AMQ48986 USA 2016/02/02)
FLR(AMM39804 Colombia 2015/12)
COL/FCC00093/2015(ANS60026 Colombia 2015/10/07)
MEX I 7(ANF04750 Mexico 2015/11)
Aedes.sp/MEX/MEX 2-81/2016(ANN83272 Mexico 2016/01/01)
103451(ANG09399 Honduras 2016/01/06)
8375(AMC13913 Guatemala 2015/11/01)
GDZ16001(AML82110 China 2016/02/12)
R114916(AOG18295 Dominican Republic 2016/06/06)
R116265(AOG18296 Mexico 2016/06/23)
17271(AML81028 French Guiana 2015/12)
Brazil-ZKV2015(AMD16557 Brazil 2015/11/30)
H.sapiens-tc/PHL/2012/CPC-0740(AMD61711 Philippines 2012/05/09)
Zika(ACD75819 Micronesia 2007/06/01)
H.sapiens-tc/THA/2014/SV0127- 14(AMD61710 Thailand 2014/07/19)
Z16006(AMR39830 China 2016/02/16)
Homo sapiens/THA/PLCal ZV/2013(AOC50652 Thailand 2013/02/12)
10015 PF(ANO46303 French Polynesia 2014/01)
FSS13025(AFD30972 Cambodia 2010)
Bahia07(AMQ76465 Brazil 2015/07/01)
P6-740(AEN75264 Malaysia 1966)
P6-740(ANK57896 Malaysia 1966/07)
MR 766(ABI54475 Uganda 1947)
Homo sapiens/NGA/IbH-30656 SM21V1-V3/1968(AMR68906 Nigeria 1968/09/09)
ARB15076(AHF49784 Central African Republic 1980)
ArB1362(AHL43500 Central African Republic 1968)
ARB13565(AHF49783 Central African Republic 1976)
ArD 41519(AEN75266 Senegal 1984)
A.africanus-tc/SEN/1984/41525-DAK(AMR39832 Senegal 1984/11/20)
MR 766(AAV34151 Uganda 1947)
MR766-NIID(BAP47441 Uganda 1947)
ArD7117(AHL43501 Senegal 1968)
ArD128000(AHL43502 Senegal 1997)
ArD157995(AHL43503 Senegal 2001)
ArD158084(AHL43504 Senegal 2001)
Asian
African 2
African 1
5

FIG. 2. Phylogenetic tree of 44 Zika viruses with NS5 proteins based on the 60-dimensional natural vector method. The strains are divided into two lineages: the African 1 and African 2 lineages. Names, accession number, location, and year identified of the strains are displayed.

One MR766 strain and another strain from Nigeria are placed together with the Asian clade. The two African strains and those in the Asian clade are grouped together as a new clade, which is a sister group to other African ZIKV viruses. Thus, we conclude that the Asian clade is originated from Africa. The appearance of MR766 strains in the two African lineages shows their diversity at the beginning of isolation. Compared with the polyprotein phylogeny shown in Figure 1, the Zika strains in Africa are not divided into the West African and East African genotypes in Figure 2. In addition, MR 766 (ABI54475, Uganda 1947) and *Homo sapiens*/NGA/IbH-30656 SM21 V1–V3/1968 (AMR68906, Nigeria 1968/09/09) are near to the Asian group in the NS5 phylogenetic tree.

Applying the natural vector method, the phylogenetic tree of 87 complete coding regions of Zika strains is constructed. The tree is shown in Supplementary Figure S1. As illustrated in the figure, these strains are divided into the Asian group (in blue) and the African group (in red). The strain *Homo sapiens*/NGA/IbH-30656 SM21 V1-V3/1968 is closest to *Aedes africanus*/SEN/DakAr41524/1984 at distance 37.38. Compared with the polyprotein tree (Fig. 1), the West African strains are not separate from the East African strains. Using the NS5 gene, the phylogeny of 44 Zika strains based on natural vectors is shown in Supplementary Figure S2. It is obvious that the strains are clustered into the Asian clade (in green) and the African clade (in red). These two clades seem parallel. However, the NS5 protein phylogeny (Fig. 2) based on natural vectors indicates that the Asian strains originate from Africa. As a result, using the natural vector method, the polyprotein tree and the NS5 tree, both provide more insights into the origin and evolution of Zika strains than those based on DNA.

Phylogeny of flaviviruses

Using the natural vector method, the evolutionary history of 53 polyprotein sequences of flaviviruses is constructed as shown in Figure 3. The phylogenetic analysis suggests that the mosquito/vertebrate viruses form two separate lineages that are phylogenetically distant. ZIKV is placed together with Spondweni virus in one of the mosquito/vertebrate groups. The two dual host-affiliated insect-specific viruses, Donggang virus and Chaoyang virus, are positioned between the two lineages. The classical insect-specific flaviviruses are clustered forming one clade that includes *Aedes* flavivirus, Kamiti River virus, Cell fusing agent virus, Mercadeo virus, Parramatta River virus, *Culex* flavivirus, Quang Binh virus, and mosquito flavivirus.

The flaviviruses only infecting vertebrates with no known vector are separated into two clades. One clade is closely associated with the mosquito/vertebrate lineages. Jutiapa virus, Yokose virus, Sokoluk virus, and Entebbe bat virus belong to this clade. The other clade is distinctly different from other viruses and constitutes Apoi virus, Modoc virus, Tamana bat virus, Rio Bravo virus, and Montana Myotis leukoencephalitis virus. The viruses transmitted by ticks are grouped to form the tick-borne lineage consisting of Kama virus, Tyuleniy virus, Alkhurma virus, Powassan virus, Karshi virus, Omsk hemorrhagic fever virus, Langat virus, tick-borne encephalitis virus, Spanish goat encephalitis

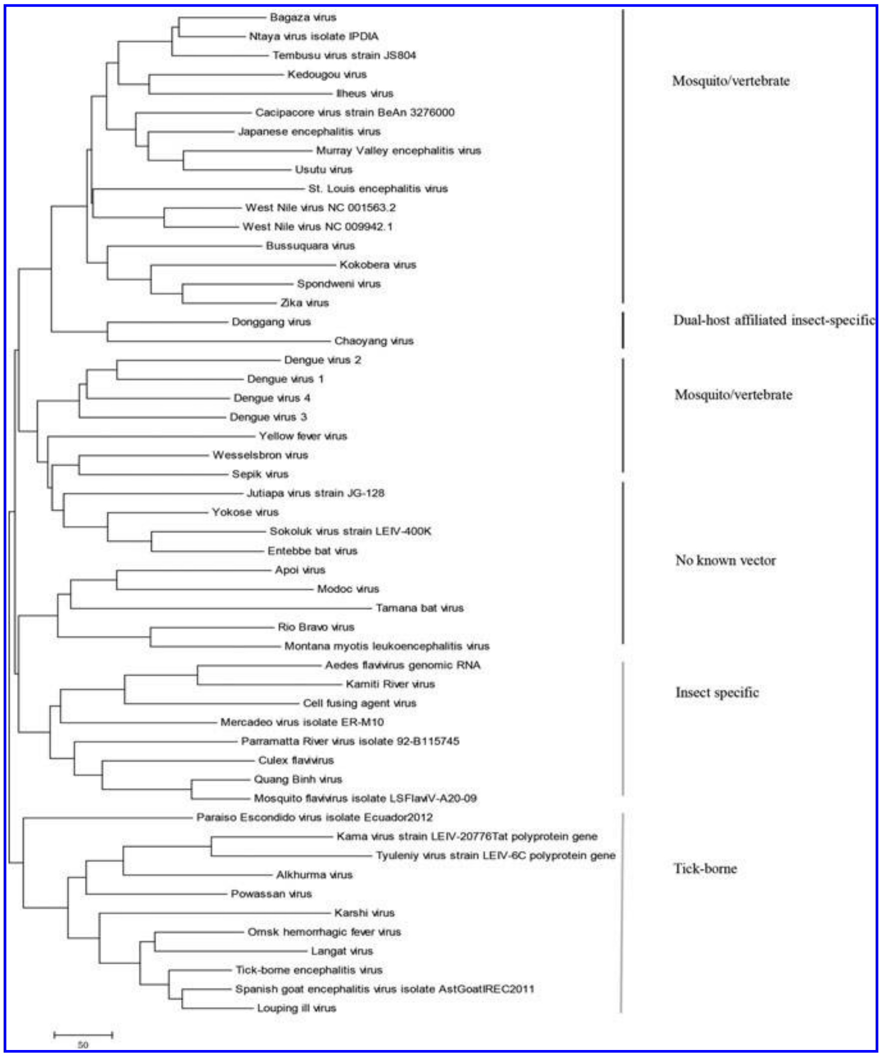
Bagaza virus
Ntaya virus isolate IPDIA
Tembusu virus strain JS804
Kedougou virus
Ilheus virus
Cacipacore virus strain BeAn 3276000
Japanese encephalitis virus
Murray Valley encephalitis virus
Usutu virus
St. Louis encephalitis virus
West Nile virus NC 001563.2
West Nile virus NC 009942.1
Bussuquara virus
Kokobera virus
Spondweni virus
Zika virus
Donggang virus
Chaoyang virus
Dengue virus 2
Dengue virus 1
Dengue virus 4
Dengue virus 3
Yellow fever virus
Wesselsbron virus
Sepik virus
Jutiapa virus strain JG-128
Yokose virus
Sokoluk virus strain LEIV-400K
Entebbe bat virus
Apoi virus
Modoc virus
Tamana bat virus
Rio Bravo virus
Montana myotis leukoencephalitis virus
Aedes flavivirus genomic RNA
Kamiti River virus
Cell fusing agent virus
Mercadeo virus isolate ER-M10
Parramatta River virus isolate 92-B115745
Culex flavivirus
Quang Binh virus
Mosquito flavivirus isolate LSFlaviV-A20-09
Paraiso Escondido virus isolate Ecuador2012
Kama virus strain LEIV-20776Tat polyprotein gene
Tyuleniy virus strain LEIV-6C polyprotein gene
Alkhurma virus
Powassan virus
Karshi virus
Omsk hemorrhagic fever virus
Langat virus
Tick-borne encephalitis virus
Spanish goat encephalitis virus isolate AstGoatIREC2011
Louping ill virus
Mosquito/vertebrate
Dual-host affiliated insect-specific
Mosquito/vertebrate
No known vector
Insect specific
Tick-borne
50

FIG. 3. Phylogenetic tree of 53 viruses in the genus Flavivirus with polyproteins based on the 60-dimensional natural vector method.

virus, and Louping ill virus. These results are consistent with the previous work (Huhtamo *et al.*, 2014; Blitvich and Firth, 2015). The unclassified Paraiso Escondido virus is near to viruses in the tick-borne clade, and thus is predicted to belong to this group.

Discussion

In this work, we use the efficient natural vector method to infer the phylogeny of 87 ZIKV strains. This method is alignment-free and does not assume any parameters. Based on translated polyprotein sequences combined with the 60-dimensional natural vector, we classify the ZIKV viruses into three lineages: Asian, East African, and West African. The tree suggests that ZIKV evolves into three independent genotypes. It also indicates that the recent ZIKV infections in the Americas are imported from Asia and Pacific regions. Based on our phylogenetic analysis using NS5 protein in Figure 2, we first find that the ZIKV strains are evolving into two African lineages. This tree implies that the newly isolated strains in Americas originate from those in the Asian and Pacific islands. Our result demonstrates that ZIKV originates from Africa and expands into Asia, Pacific islands, and Americas. The MR766 strain (accession number: ABI54475) is one representative ancestral genotype for the Asian clade. Thus, we provide a novel evidence supporting the African origin hypothesis of ZIKV. This finding will help to track circulation of ZIKV in the world and elucidate the evolutionary history of ZIKV.

In Figure 1, the *H. sapiens*-tc/PHL/2012/CPC-0740 (AMD61711, Philippines) is nearest to the Asian genotype *Aedes* sp./MYS/P6-740/1966 (ANW07475) isolated in Malaysia in 1966. In Figure 2, however, the Philippines strain is closest to the Zika strain (ACD75819) isolated in Micronesia in 2007. The different position of the Philippines strain in two trees may be caused by the difference of mutation rate between the polyprotein and NS5 protein. The NS5 protein possibly has undergone higher mutation rate than the polyprotein. In addition, we also perform the phylogenetic analysis for the 53 viruses in the genus *Flavivirus*. The viruses are well classified into five groups. Each group is associated with typical transmission mode and host range. In particular, ZIKV is genetically near to Spondweni virus and both of them are positioned in one of the mosquito /vertebrate groups.

In the work of Lanciotti *et al.* (2016), three major lineages, Asian, West African, and East African lineages, are characterized. The result is based on the complete genomes and alignment. However, the number of ZIKV available

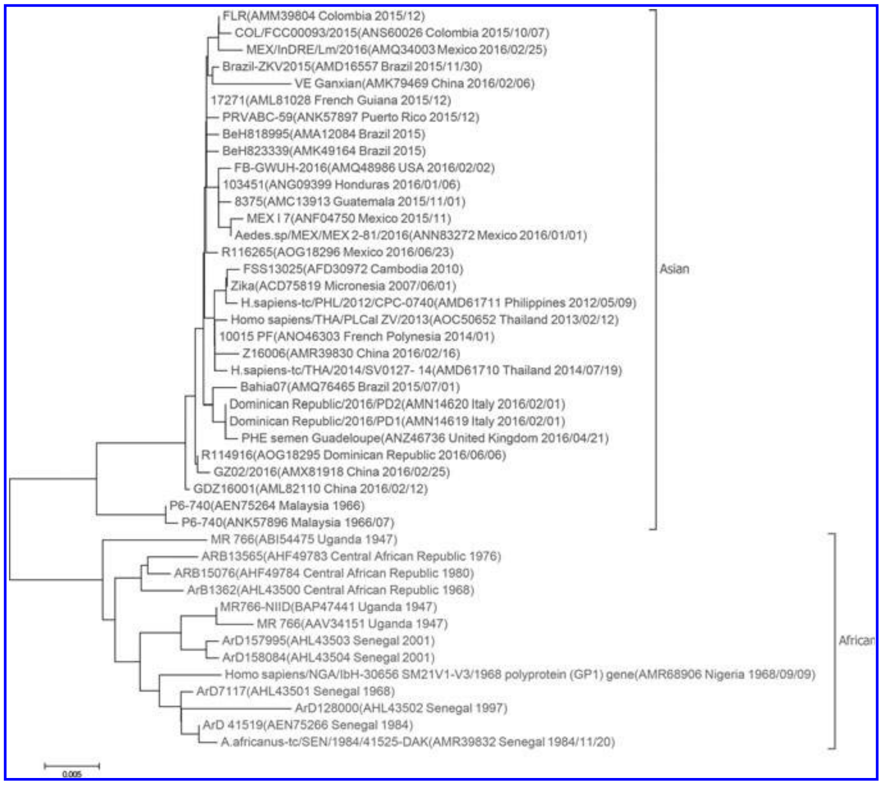
FLR(AMM39804 Colombia 2015/12)
COL/FCC00093/2015(ANS60026 Colombia 2015/10/07)
MEX/InDRE/Lm/2016(AMQ34003 Mexico 2016/02/25)
Brazil-ZKV2015(AMD16557 Brazil 2015/11/30)
VE Ganxian(AMK79469 China 2016/02/06)
17271(AML81028 French Guiana 2015/12)
PRVABC-59(ANK57897 Puerto Rico 2015/12)
BeH818995(AMA12084 Brazil 2015)
BeH823339(AMK49164 Brazil 2015)
FB-GWUH-2016(AMQ48986 USA 2016/02/02)
103451(ANG09399 Honduras 2016/01/06)
8375(AMC13913 Guatemala 2015/11/01)
MEX I 7(ANF04750 Mexico 2015/11)
Aedes.sp/MEX/MEX 2-81/2016(ANN83272 Mexico 2016/01/01)
R116265(AOG18296 Mexico 2016/06/23)
FSS13025(AFD30972 Cambodia 2010)
Zika(ACD75819 Micronesia 2007/06/01)
H.sapiens-tc/PHL/2012/CPC-0740(AMD61711 Philippines 2012/05/09)
Homo sapiens/THA/PLCal ZV/2013(AOC50652 Thailand 2013/02/12)
10015 PF(ANO46303 French Polynesia 2014/01)
Z16006(AMR39830 China 2016/02/16)
H.sapiens-tc/THA/2014/SV0127- 14(AMD61710 Thailand 2014/07/19)
Bahia07(AMQ76465 Brazil 2015/07/01)
Dominican Republic/2016/PD2(AMN14620 Italy 2016/02/01)
Dominican Republic/2016/PD1(AMN14619 Italy 2016/02/01)
PHE semen Guadeloupe(ANZ46736 United Kingdom 2016/04/21)
R114916(AOG18295 Dominican Republic 2016/06/06)
GZ02/2016(AMX81918 China 2016/02/25)
GDZ16001(AML82110 China 2016/02/12)
P6-740(AEN75264 Malaysia 1966)
P6-740(ANK57896 Malaysia 1966/07)
Asian
MR 766(ABI54475 Uganda 1947)
ARB13565(AHF49783 Central African Republic 1976)
ARB15076(AHF49784 Central African Republic 1980)
ArB1362(AHL43500 Central African Republic 1968)
MR766-NIID(BAP47441 Uganda 1947)
MR 766(AAV34151 Uganda 1947)
ArD157995(AHL43503 Senegal 2001)
ArD158084(AHL43504 Senegal 2001)
Homo sapiens/NGA/IbH-30656 SM21V1-V3/1968 polyprotein (GP1) gene(AMR68906 Nigeria 1968/09/09)
ArD7117(AHL43501 Senegal 1968)
ArD128000(AHL43502 Senegal 1997)
ArD 41519(AEN75266 Senegal 1984)
A.africanus-tc/SEN/1984/41525-DAK(AMR39832 Senegal 1984/11/20)
African
0.005

FIG. 4. Phylogenetic tree of 44 Zika viruses with NS5 proteins based on multiple sequence alignment. The alignment is obtained by ClustalW and the tree is derived by neighbor-joining algorithm incorporated in MEGA 6.0.

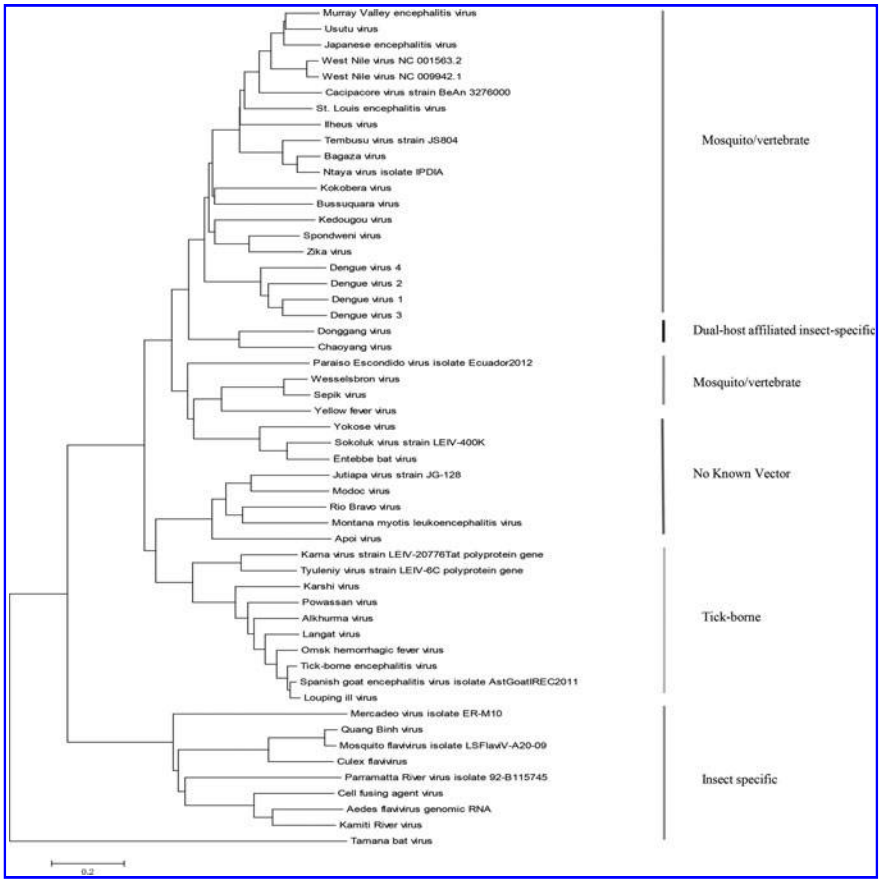
Murray Valley encephalitis virus
Usutu virus
Japanese encephalitis virus
West Nile virus NC 001563.2
West Nile virus NC 009942.1
Cacipacore virus strain BeAn 3276000
St. Louis encephalitis virus
Ilheus virus
Tembusu virus strain JS804
Bagaza virus
Ntaya virus isolate IPDIA
Kokobera virus
Bussuquara virus
Kedougou virus
Spondweni virus
Zika virus
Dengue virus 4
Dengue virus 2
Dengue virus 1
Dengue virus 3
Donggang virus
Chaoyang virus
Paraiso Escondido virus isolate Ecuador2012
Wesselsbron virus
Sepik virus
Yellow fever virus
Yokose virus
Sokoluk virus strain LEIV-400K
Entebbe bat virus
Jutiapa virus strain JG-128
Modoc virus
Rio Bravo virus
Montana myotis leukoencephalitis virus
Apoi virus
Karna virus strain LEIV-20776Tat polyprotein gene
Tyuleniy virus strain LEIV-6C polyprotein gene
Karshi virus
Powassan virus
Alkhurma virus
Langat virus
Omsk hemorrhagic fever virus
Tick-borne encephalitis virus
Spanish goat encephalitis virus isolate AstGoatIREC2011
Louping ill virus
Mercadeo virus isolate ER-M10
Quang Binh virus
Mosquito flavivirus isolate LSFlaviV-A20-09
Culex flavivirus
Parramatta River virus isolate 92-B115745
Cell fusing agent virus
Aedes flavivirus genomic RNA
Kamiti River virus
Tamana bat virus
Mosquito/vertebrate
Dual-host affiliated insect-specific
Mosquito/vertebrate
No Known Vector
Tick-borne
Insect specific
0.2

FIG. 5. Phylogenetic tree of 53 viruses in the genus Flavivirus with polyproteins based on multiple sequence alignment. The alignment is obtained by ClustalW and the tree is derived by neighbor-joining algorithm incorporated in MEGA 6.0.

used in their analysis is limited. The protein sequences of ZIKV are more conserved than DNA sequences and play important roles in viral activities. It may serve as markers to infer ZIKV evolution. Using complete protein sequences, the Zika strains are clustered into three clades, which is consistent with their result. In addition, all available ZIKV strains with full-length polyproteins are included in our analysis, which shows more diversity among the ZIKV strains.

In comparison, the phylogenetic tree based on alignment of NS5 proteins is built. The tree is obtained by ClustalW and the neighbor-joining algorithm implemented in MEGA 6.0. As shown in Figure 4, two major lineages, Asian and African lineages, are identified. However, the divergence time for the lineages is not clear. Using NS5 proteins, our phylogenetic tree construction by the natural vector method implies that ZIKV strains in Africa diverge earlier from the common ancestral than those in Asian clade. Compared with alignment-based methods using complete or partial genomes of ZIKV, our natural vector method using protein sequences provides novel insights into the origin, divergence, and evolution of ZIKV. Using the polyprotein sequences and multiple sequence alignment, a neighbor-joining phylogenetic tree of 53 flaviviruses is also plotted in Figure 5. Five clades are formed, which is in accordance with the tree derived from the natural vector method.

Various numerical representations of genetic sequences have been developed in the past decades. Recently, a two-dimensional (2D) graphical representation method is utilized to characterize Zika viruses (Nandy *et al.*, 2016). Using the method, a genomic sequence of ZIKV is mapped to a 2D curve. For the curve, two parameters are defined to represent its center. The distance from the center to the origin denoting by g serves as sequence descriptor. However, the curve may have degeneracy and loops as well. The relationship between the sequence and the curve is not one-to-one. Moreover, the sequence descriptor may lose a lot of information of the curve. For example, MR 766 (HQ234498, Uganda 1947) has $g = 84.13063$ and its nearest strain is MEX I 7 (KX247632, Mexico 2015/11), where $g = 84.13051$. However, the two strains are from African and Asian genotypes, respectively. By our natural vector method, their distance is 69.404 and the nearest neighbor of MR 766 (HQ234498, Uganda 1947) is MR-766 (KX377335, Uganda 1947/04) with distance 5.459111.

Our numerical representation using 60-dimensional vectors is an efficient dimensionality reduction method for a long protein sequence. For each amino acid A, let n_A be the number of A, μ_A be the average, and D_2^A be the variation of positions of A in the sequence. These numbers can characterize the distribution of A in the sequence and display the differences between protein sequences by plotting $\left(n_A, \mu_A, D_2^A\right)$ in three-dimensional space. Unlike multiple sequence alignment methods, once the natural vectors of viruses are computed, they can be stored in a database. To compare genomes, for a new sequence, we just need to compute its own natural vector and compare it with the database.

One of the distinct advantages of natural vector is its efficiency. It is fast to convert sequences to numerical vectors and does not presume any model to infer phylogeny. The previous studies for ZIKV phylogeny are predominately generated by multiple sequence alignment that is very time consuming in computation. With the ClustalW algorithm implemented in MEGA 6.0 software, it takes 6 min to align the 44 NS5 protein sequences in a laptop with 8 GB RAM, Intel(R) Core(TM) i7-4500U CPU, and 1.80 GHz, while it only takes 0.26 s to convert these sequences to the natural vectors. Moreover, it needs 80 min to accomplish alignment for the 53 flaviviruses implemented by ClustalW, while the natural vector method only requires 1.09 s. Using our natural vector method, the classification of 776 single-segmented dsDNA viruses in the RefSeq database may be completed within 76.7 min on a PC computer (CPU 1.67 GHz, 3 GB of RAM). For multiple sequence alignment, this task is very difficult even infeasible in computation (Yu *et al.*, 2013). According to the comparison using several examples, we conclude that the natural vector approach is a powerful tool to perform phylogeny analysis. It is accurate and competitive to alignment-based methods. It also has advantage over alignment-based methods in computational time.

Conclusions

In this study, we employ an alignment-free method, natural vector, to explore the evolutionary relationship among 87 ZIKV strains for the first time. The protein sequences of ZIKV are first introduced for phylogeny analysis. Based on the polyproteins, three major lineages are identified: the West African, the East African, and the Asian lineages. Applying the NS5 proteins, we find that the strains are divided into two African clades. This result confirms the hypothesis that ZIKV originates from Africa. Both the polyprotein tree and NS5 protein tree show that the ZIKV epidemic in Americas comes from Asia and Pacific islands. Our findings help to elucidate the origin and evolution of ZIKV. Besides, the phylogeny of 53 flaviviruses, including ZIKV, is constructed by the natural vector method based on polyproteins. In the tree, mosquito/vertebrate, dual host-affiliated insect-specific, classical insect-specific flaviviruses, tick-borne, and the no known vector groups are characterized. Especially, ZIKV is close to Spondweni virus belonging to the mosquito/vertebrate group. Our phylogenetic analysis of ZIKV and other flaviviruses may be helpful for the surveillance of these viruses and the vaccine design.

Acknowledgments

This study is supported by the National Natural Sciences Foundation of China (31271408 to S.S.-T.Y.), Tsinghua University startup fund (to S. S.-T.Y.), and Tsinghua University Independent Research Project Grant (to S. S.-T.Y.). The funders did not take part in study design, in collection and analysis of data, in the writing of the article, or in the decision to publish this article.

Disclosure Statement

No competing financial interests exist.

References

Baronti, C.E.C., *et al.* (2014). Complete coding sequence of Zika virus from a French polynesia outbreak in 2013. Genome Announc **2,** e00500–e00514.

Billoir, F., *et al.* (2000). Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector. J Gen Virol **81,** 781–790.

Blitvich, B.J., and Firth, A.E. (2015). Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. Viruses **7,** 1927–1959.

Broutet, N., *et al.* (2016). Zika virus as a cause of neurologic disorders. N Engl J Med **374,** 1506–1509.

Cao-Lormeau, V.-M., *et al.* (2014). Zika virus, French polynesia, South pacific, 2013. Emerg Infect Dis **20,** 1085–1086.

Dai, Q., *et al.* (2008). Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics **24,** 2296–2302.

Deng, M., *et al.* (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. PloS One **6,** e17293.

Duffy, M.R., *et al.* (2009). Zika virus outbreak on Yap Island, Federated States of Micronesia. N Engl J Med **360,** 2536–2543.

Enfissi, A., *et al.* (2016). Zika virus genome from the Americas. Lancet **387,** 227–228.

Faria, N.R., *et al.* (2016). Zika virus in the Americas: early epidemiological and genetic findings. Science **352,** 345–349.

Faye, O., *et al.* (2014). Molecular evolution of Zika virus during its emergence in the 20(th) century. PLoS Negl Trop Dis **8,** e2636.

Gong, Z., *et al.* (2016). Zika virus: two or three lineages? Trends Microbiol **24,** 521–522.

Huang, H.-H., *et al.* (2014). Global comparison of multiple-segmented viruses in 12-dimensional genome space. Mol Phylogenet Evol **81,** 29–36.

Huhtamo, E., *et al.* (2014). Novel flaviviruses from mosquitoes: mosquito-specific evolutionary lineages within the phylogenetic group of mosquito-borne flaviviruses. Virology **464–465,** 320–329.

Lanciotti, R.S., *et al.* (2008). Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. Emerg Infect Dis **14,** 1232–1239.

Lanciotti, R.S., *et al.* (2016). Phylogeny of Zika virus in Western hemisphere, 2015. Emerg Infect Dis **22,** 933–935.

Li, Y., *et al.* (2016). Virus classification in 60-dimensional protein space. Mol Phylogenet Evol **99,** 53–62.

Moureau, G., *et al.* (2015). New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. PLoS One **10,** e0117849.

Nandy, A., *et al.* (2016). Characterizing the Zika virus genome—a bioinformatics study. Curr Comput Aided Drug Des **12,** 87–97.

Shen, S., *et al.* (2016). Phylogenetic analysis revealed the central roles of two African countries in the evolution and worldwide spread of Zika virus. Virol Sin **31,** 118–130.

Wu, G.A., *et al.* (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proc Natl Acad Sci U S A **106,** 12826–12831.

Yu, C., *et al.* (2013). Real time classification of viruses in 12 dimensions. PloS One **8,** e64328.

Zanluca, C., *et al.* (2015). First report of autochthonous transmission of Zika virus in Brazil. Mem Inst Oswaldo Cruz **110,** 569–572.

Zhu, Z., *et al.* (2016). Comparative genomic analysis of pre-epidemic and epidemic Zika virus strains for virological factors potentially associated with the rapidly expanding epidemic. Emerg Microbes Infect **5,** e22.

Address correspondence to:
Stephen S.-T. Yau
Department of Mathematical Sciences
Tsinghua University
Beijing 100084
P.R. China

E-mail: yau@uic.edu

Received for publication September 21, 2016; received in revised form November 2, 2016; accepted November 4, 2016.