Possible Roles of New Mutations Shared by Asian and American Zika Viruses

Shozo Yokoyama, Emory University
William T. Starmer, Syracuse University

Journal Title: Molecular Biology and Evolution
Volume: Volume 34, Number 3
Publisher: Oxford University Press | 2017-01-12, Pages 525-534
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/molbev/msw270
Permanent URL: https://pid.emory.edu/ark:/25593/s2jrm

Final published version: http://dx.doi.org/10.1093/molbev/msw270

Copyright information:
© The Author 2017.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/).

Accessed August 5, 2017 2:27 AM EDT
Possible Roles of New Mutations Shared by Asian and American Zika Viruses

Shozo Yokoyama*1 and William T. Starmer2
1Department of Biology, Emory University, Atlanta, GA
2Department of Biology, Syracuse University, Syracuse, NY
*Corresponding author: E-mail: syokoya@emory.edu.
Associate editor: Takashi Gojobori

Abstract

Originating in Africa, the Zika virus (ZIKV) has spread to Asia, Pacific Islands and now to the Americas and beyond. Since the first isolation in 1947, ZIKV strains have been sampled at various times in the last 69 years, but this history has not been reflected in studying the patterns of mutation accumulation in their genomes. Implementing the viral history, we show that the ZIKV ancestor appeared sometime in 1930–1945 and, at that point, its mutation rate was probably less than 0.2 × 10⁻³/nucleotide site/year and subsequently increased significantly in most of its descendants. Sustaining a high mutation rate of 4 × 10⁻³/site/year throughout its evolution, the Ancestral Asian strain, which was sampled from a mosquito in Malaysia, accumulated 13 mutations in the 3'-untranslated region of RNA stem-loops prior to 1963, seven of which generate more stable stem-loop structures and are likely to inhibit cellular antiviral activities, including immune and RNA interference (RNAi) pathways. The seven mutations have been maintained in all Asian and American strains and may be responsible for serious medical problems we are facing today and offer testable hypotheses to examine their roles in molecular interactions during brain development.

Key words: Zika viruses, evolutionary tree, evolutionary rates, RNA structures.

Introduction

The Zika virus (ZIKV) is a member of family Flaviviridae, genus Flavivirus, and is transmitted among humans primarily by Aedes mosquitoes. The single-stranded positive-sense RNA virus has a length of approximately 10,800 nucleotides with a 5' - and 3'-UTR. It encodes a single polyprotein that undergoes post-translational processing to yield capsid (C), membrane precursor (prM), envelope (E), and seven nonstructural (NS) proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) (Chambers et al. 1990; Cunha et al. 2016). The ZIKV (MR766) was first isolated in 1947 from a sentinel monkey (Macaca mulatta) in the Zika forest of Uganda (Dick et al. 1952). Although dispersing within the African Continent, some strains also spread to Malaysia and then to Philippine and Micronesia, forming a new Asian lineage (Haddow et al. 2012; Faye et al. 2014). After sporadic human infections in Africa and Asia, ZIKV outbreaks were reported in Micronesia (Duffy et al. 2009) in 2007 and French Polynesia (Musso, Nilles, et al. 2014) in 2013–2014. It is suspected that ZIKV reached Brazil during the Polynesian epidemic (Faria et al. 2016) and has spread rapidly throughout the Americas (Chang et al. 2016; Faria et al. 2016).

About 80% of ZIKV infections in humans are asymptomatic and when they appear the symptoms are usually mild, including rash, fever, red eyes, joint pain, and headache (Duffy et al. 2009; Musso, Nhan, et al. 2014; Malone et al. 2016). This description of ZIKV infection had to be changed drastically in the summer of 2015 when the association between ZIKV and microcephaly, Guillain–Barre syndrome and other serious brain disorders was found in the Americas (Calvet et al. 2016; Mlakar et al. 2016; Schuler-Faccini et al. 2016) and retrospectively in French Polynesia (Cauchemez et al. 2016). Because of these findings and the authentication as a direct causative agent by the US Centers for Disease Control and Prevention (CDC) (Rasmussen et al. 2016), ZIKV has become an emerging infectious public threat (Calvet et al. 2016; Cauchemez et al. 2016; Chang et al. 2016; Malone et al. 2016; Mlakar et al. 2016; Rasmussen et al. 2016; Schuler-Faccini et al. 2016). Indeed, using human stem cells (Garcez et al. 2016; Tang et al. 2016) and mice (Cugola et al. 2016; Li et al. 2016; Miner et al. 2016), ZIKV is shown to target neural stem cells (neurospheres), brain organoids, placenta, fetal brains and to cause microcephaly; ZIKV has also been found in human fetal brains (Mlakar et al. 2016). Curiously, only a portion (1–13%) of mothers infected with ZIKV have babies with microcephaly and other brain dysfunctions (Cauchemez et al. 2016; Johansson et al. 2016).

To understand the molecular mechanisms, by which ZIKV causes these serious medical problems, one crucial step is to identify mutations that induce cytopathicity and pathogenicity. One effective way of predicting these mutations is to study how they have accumulated in different ZIKV strains and their possible functions. Expansions of ZIKV strains to various geographic regions and potentially critical mutations have been studied primarily by constructing phylogenetic
trees of ZIKV and other flaviviruses (Duffy et al. 2009; Haddow et al. 2012; Cao-Lormeau and Musso 2014; Faye et al. 2014; Chang et al. 2016; Enfissi et al. 2016; Faria et al. 2016; Shen et al. 2016; Zhu et al. 2016). These evolutionary trees show that 1) the strains isolated from Asia, Pacific Islands and the Americas (collectively called Asian strains) and their ancestral African strains form two evolutionarily distinct groups and 2) the numbers of nucleotide substitutions per site (or branch lengths) of various strains measured from the ZIKV ancestor are very similar. However, at least the latter result is in error because MR766 basically stopped evolving in 1947 and American strains accumulated additional mutations for the next 69 years.

As described below, the currently known evolutionary trees of different ZIKV strains have been constructed by basically assuming that various strains evolved with similar speeds (Faria et al. 2016). When this assumption is removed, drastically different tree topologies of ZIKV strains emerge. These new evolutionary trees allow us to properly evaluate the numbers of nucleotide (or amino acid) substitutions per year (or evolutionary rates of mutant substitution) for different strains and also identify potentially critical mutations in the Asian strains under the proper historical context.

The goal of this paper is to find a biologically acceptable evolutionary tree of ZIKV strains and try to identify critical mutations. For this purpose, we consider some representative ZIKV with complete genome sequences and exclude partially sequenced African strains and many completely sequenced American strains without losing generality. Throughout the paper, the strains collected from humans are denoted by the countries of their origin and those isolated from mosquitoes are indicated by abbreviated species names. The roots of these trees were determined using the Spondweni virus (SM-6V-1) as the outgroup (indicated by stars, supplementary table S1, Supplementary Material online).

**Results and Discussion**

**Conventional Phylogenetic Trees of ZIKV Strains**

Before finding, a new evolutionary tree of ZIKV strains, those of 9 Asian and 11 African ZIKV strains were constructed first by applying the maximum likelihood (ML), neighbour-joining (NJ), and maximum parsimony (MP) methods directly to their nucleotide sequences using the MEGA6 web server (Tamura et al. 2013). The resulting three evolutionary trees show virtually identical topologies and the NJ and ML trees have similar branch lengths (fig. 1).

**Fig. 1.** Phylogenetic trees of 20 representative ZIKV strains. Numbers in front of ZIKV strains indicate the year of isolation. The root was determined using Spondweni virus (SM-6V-1) as the outgroup. (A) The NJ tree, which was obtained by applying the Neighbour-Joining (NJ)-method directly to the sequence data (Tamura et al. 2013). Numbers at various nodes indicate clustering percent supports generated by 1000 bootstrap analyses. Blue and orange rectangles indicate groups 1–3 of African strains and Asian strains which were distinguished using the NJ method, respectively. (B) The PE tree, where the tree topology was determined first and then branch lengths were determined using the ML method (Yang 2007).
differences: MR766 and Ada1 switched their phylogenetic positions in group 1, whereas Alu2 and Aaf4 switched their phylogenetic positions in group 3.

Another phylogenetic tree was obtained using the Bayesian Evolutionary Analyses Sampling Trees (BEAST v1.8.4) method (Drummond and Rambaut 2007). Running $100 \times 10^6$ generations with the substitution model (HKY + three codon positions), the clock model (strict) and the tree model (constant), the minimum effective sample size (ESS) of 1795 was obtained (supplementary fig. S1C, Supplementary Material online). Despite assimilating the “time-stamped data,” which certainly show a shorter branch length for MR766, the BEAST tree has a very similar topology as those of the ML, MP, and NJ trees (supplementary fig. S1C, Supplementary Material online). The BEAST trees based on the uncorrelated lognormal clock ($300 \times 10^6$ generations, ESS $> 376$) and the uncorrelated exponential clock ($300 \times 10^6$ generations, ESS $> 326$) are very similar to the BEAST tree with the strict clock model, except that Alu2 and Nigeria are switched within group 3.

All of these trees show that the ancestor of African strains MR766, Ada1, and Ada2 (group 1) split from the common ancestor of Aaf1, Aaf2, Aaf3, and Aoep (group 2) and Alu1, Alu2, Aaf4, and Nigeria (group 3). These groupings are depicted in the blue rectangle (fig. 1A). Supported by high bootstrap values (the NJ, ML, and MP trees) and high posterior probabilities (the BEAST tree), these evolutionary steps are, statistically speaking, highly reliable and show that MR766 appeared only recently, which is consistent with the currently known evolutionary trees of ZIKV strains (Haddow et al. 2012; Fay et al. 2014; Enfissi et al. 2016; Faria et al. 2016; Shen et al. 2016; Zhu et al. 2016). In reality, however, MR766 was isolated in 1947 and has a low passage history (Haddow et al. 2012), it is expected to have a much shorter branch length than many other African strains.

To explain these “conventional” tree topologies and similar branch lengths of different strains measured from the ZIKV ancestor, we have to assume that the evolutionary rates of nucleotide substitution (or mutation rates) had to be rigorously adjusted among them so that their branch lengths become very similar at various times of virus isolation—a highly improbable proposition.

A New Evolutionary Tree

To implement the variable sampling history and possibly different evolutionary rates of ZIKV strains, a rooted phylogenetic tree of the same 20 ZIKV strains was constructed first by obtaining its topology based on the shared amino acid replacements (or parallel changes) in their entire polyproteins (supplementary figs. S2 and S3, Supplementary Material online) and then evaluating the numbers of nucleotide substitutions per site for all branches using the ML-method (Yang 2007). In the following, this new evolutionary tree of ZIKV is referred to simply as the parallel evolution (PE) tree (fig. 18).

The PE tree differs dramatically from the traditional trees (fig. 1A, supplementary fig. S1A–C, Supplementary Material online) and exhibits three major characteristics. First, as expected, MR766 is most distantly related to the other ZIKV strains and consequently the African and Asian strains cannot be clearly distinguished; moreover, MR766 has a very short branch length after it split from the common ancestor of the other ZIKV strains. Second, the ancestors of four African lineages (MR766, Ada1, Ada2, and the rest) and Aae-Malaysia diverged from each other about the same time, strongly suggesting that ZIKV evolution experienced a bottleneck. Third, the branch lengths of different strains, exemplified by Aae-Malaysia and Ada2, which were isolated in 1966 and 2001, respectively, do not reflect the times of their isolation, showing that the evolutionary rates of nucleotide substitution are extremely variable among different strains. However, the PE tree and the conventional trees have one common feature—the American strains can be traced back to the ancestors of Polynesia, Thailand, Cambodia, Micronesia (or Philippine) and Aae-Malaysia, in that order.

In constructing the PE tree, it was assumed that the ancestral Asian strain diverged from the common ancestor of Ada1 and Ada2 because of one shared amino acid replacement at position 3040 in the polyprotein (supplementary fig. S2, Supplementary Material online). Recently, considering the E and NSS regions separately, it has been claimed that ZIKV strains should be distinguished into two distinct lineages—lineage 1 includes the African strains we are considering (fig. 1A and B), whereas lineage 2 contains ZIKV strains sampled from Senegal and Cote d’Ivoire (Shen et al. 2016). Again, the trees based on the E and NSS regions show that MR766 appeared only recently and are unrealistic. However, comparing these and the ZIKV sequences considered here (supplementary table S1, Supplementary Material online), we can identify four shared amino acid changes: mutations from I to V (I $\rightarrow$ V) (at position 3035 in the polyprotein), N $\rightarrow$ S (at 3040) and R $\rightarrow$ Q (at 3041) that are shared by Ada1, Ada2, lineage 2 and Asian strains and also Q $\rightarrow$ H (at 3150) is shared by lineage 2 and African group 3. Hence, the Asian strains seem to have evolved from the common ancestor of Ada1, Ada2, and the lineage 2 strains. To check the effect of this assumption on the final PE tree, another tree was obtained by assuming that the Asian and African ancestors took independent paths after splitting from MR766. This alternate tree (supplementary fig. S4A, Supplementary Material online) is practically indistinguishable from the PE tree (fig. 18) and the three major characteristics of ZIKV evolution still hold.

The conventional trees, represented by the NJ tree (fig. 1A), can be easily explained from the PE tree (fig. 18) by considering the highly variable evolutionary rates of ZIKV strains and the methods by which phylogenetic trees are constructed. After all, these trees are constructed basically by searching for strains with smallest sequence differences and clustered together sequentially. Hence, because of their shortest branch lengths measured from the ZIKV ancestor (fig. 1B), MR766, Ada1, and Ada2 appear to belong to the same evolutionary group in the conventional tree (group 1). By the same token, groups 2 and 3 form separate groups and group 1 seem to be more closely related to group 2 than to group 3. Finally, because of the extremely long branch length of Aae-Malaysia, the three groups of African strains and Asian strains seem to form two distinct evolutionary groups in the conventional trees.
Mutation Rates
To evaluate the evolutionary rate of nucleotide substitution of a ZIKV strain, we need to know the duration between its appearance and the time of sampling (or divergence time). Since the branch length of MR766 is virtually zero (4 × 10⁻⁶/site), the PE tree (fig. 1B) suggests that the ZIKV ancestor appeared in 1947, the year of its sampling. However, as described below, MR766 also accumulated some mutations and the PE tree, particularly earlier stages of ZIKV evolution, must be interpreted with caution. In fact, when another tree was constructed considering the numbers of nucleotide substitution per “codon” (Yang 2007), the branch lengths of MR766 and the common ancestor of the rest of ZIKV strains are about the same (1 × 10⁻⁶/codon, supplementary fig. S4B, Supplementary Material online). Hence, it is more likely that the Asian and African strains, MR766 excluded, started to differentiate at around 1947 and the ZIKV ancestor appeared earlier than that (see Origin of ZIKV).

Divergence times (T1−T3) and evolutionary rates of nucleotide substitution (α1−α19) of ZIKV strains were evaluated by assuming that Ada1, Ada2, Aaf1, and Aae-Malaysia diverged from each other in April of 1947 and setting several short branch lengths to zero (fig. 2A). The initial evolutionary steps of the Asian strains were studied by considering Aae-Malaysia and Micronesia together with MR766 and Aaf1 (fig. 2B). Micronesia, rather than Philippine, was used here because their ancestors diverged from Aae-Malaysia at about the same time, but Micronesia was sampled five years earlier than Philippine (supplementary table S1, Supplementary Material online).

Using 10,176 comparable nucleotide sites, the branch length for Aae-Malaysia (L1) was evaluated to be 0.081/site from (d₀,1 − d₀,12 + d₁,12)/2, where d₀,i,j are the branch lengths between strains i and j (i, j = MR766 [strain 0], Aae-Malaysia [strain 1], and Aaf1 [strain 12]) (fig. 2B, supplementary table S1, Supplementary Material online). These nucleotide substitutions occurred during the period between April 1947 and July 1966, or 19.25 years, and therefore α1 = L1/19.25 = 4.2 ± 0.15 × 10⁻³/site/year. Next, the branch lengths between node O and Micronesia (L3) and between node O and Malaysia (L1,1) were estimated to be 0.030/site and 0.013/site, respectively, and T1 = 19.25 × (L1 − L1,1)/L1 = 16.2 years and α₄ = L3/(60.17 − T₁) = 0.7 ± 0.04 × 10⁻³/site/year (see Materials and Methods). Similarly, using appropriate formulae (supplementary table S2, Supplementary Material online), T2−T₆ and the remaining evolutionary rates were evaluated.

The results reveal five characteristics of ZIKV evolution. First of all, the divergence times for the 5 major evolutionary steps of Asian strains vary between 4 and 32 years, whereas those for the 4 major steps of groups 2 and 3 of African strains range between 4 and 12 years (table 1). Hence, it is likely that the ancestor of the currently known American strains split from the ancestors of Polynesia, Thailand, Cambodia, Micronesia (or Philippine), and Aae-Malaysia roughly in 2013, 2007, 1999, 1995, and 1963, respectively. In African group 2 strains, Aop appeared roughly in 1961 and Afa2 and Aaf3 in 1973, whereas Alu1 split from the ancestor of Aaf1 in 1957 and the other members of group 3 appeared in 1965.

Second, excluding MR766, the 19 strains can be classified into three groups based on their evolutionary rates: 1) Ada2 (0.05 × 10⁻³/site/year), 2) Aae-Malaysia, Aaf1, Aaf2, Alu1, and Nigeria (2.1−4.4 × 10⁻³/site/year), and 3) the remaining 13 strains (0.2−1.5 × 10⁻³/site/year) (table 2, fig. 2A). Hence, the evolutionary rates of ZIKV strains differ by two orders of magnitudes and are far more variable than previously thought (Faria et al. 2016). The extremely variable evolutionary rates among different strains seem to be the major reason why the BEAST tree with the “time-stamped data” results in a typical traditional tree. Third, as expected from the functional constraints of protein-coding genes in DNA and RNA genomes (Gojobori and Yokoyama 1985; Nei 2013), the evolutionary rate tends to be highest at the third position of a codon, followed by the first and second positions of a codon, in that order (table 2). Fourth, in the rapidly changed Aae-Malaysia, mutations occurred uniformly throughout the protein-coding regions; for example, the largest difference between the evolutionary rates of capsid (3.4 × 10⁻³/site/year) and propeptide (6.0 × 10⁻³/site/year) segments do not differ significantly. However, the evolutionary rates of amino acid replacement for NS2B and NS4A (~0.4 × 10⁻³/site/year) and capsid (2.7 × 10⁻³/site/year) differ significantly (P < 0.05) and the evolutionary rate for E (0.8 × 10⁻³/site/year) does not differ significantly from the average rate for the polyprotein (1.0 × 10⁻³/site/year) (supplementary table S3, Supplementary Material online). These characteristics differ significantly from human immunodeficiency viruses (HIV), in which envelope genes (~1.0 × 10⁻²/site/year) evolved much faster than the rest of the genome (~1.0 × 10⁻³/site/year) (Li et al. 1988; Yokoyama 1991). Finally, Brazil-2 suggests that the evolutionary rates of nucleotide substitution at the 5' and 3'-UTR are statistically indistinguishable from those of the coding region (supplementary table S3, Supplementary Material online).

Origin of ZIKV
To study when the ancestors of various ZIKV strains appeared, the nucleotide sequences of MR766 (strain 0) and Spondweni virus (strain S7) were compared with a designated third strain and the length of the third strain-specific branch (L₃rd) was evaluated. For each third strain, the time span between the year of its divergence from MR766 (the time of its appearance) and the year of its isolation in nature was determined by dividing L₃rd values by their evolutionary rates (α₃rd) and then the year of its actual appearance was estimated by subtracting this L₃rd/α₃rd value from the year it was sampled. With the exception of Ada1, the branch lengths (L₃rd) and the resulting estimates for the origin of ZIKV are not affected much by using either the gamma distance with the JC model or the standard JC model (supplementary table S4, Supplementary Material online). The results show that Ada2 with the slowest evolutionary rate (α₃rd = 0.05 × 10⁻³/site/year) has a L₃rd/α₃rd of 56 years, implying that the Ada2 ancestor appeared in 1945 (supplementary table S4, Supplementary Material online). It also
suggests that the ancestor of Ada1 appeared in 1928–1941. On the other hand, Nigeria with the fastest evolutionary rate \( (a_{3rd} = 4.4 \times 10^{-3} / \text{site/year}) \) has \( L_{3rd}/a_{3rd} \) of 8.8 years and suggests that its ancestor appeared in 1959. Likewise, the evolutionary rates for Aae-Malaysia, Aaf1, and Aop lineages suggest that the ZIKV ancestor appeared sometimes during 1952–1958. Since the ZIKV ancestor appeared prior to the isolation of MR766, the latter results are unrealistic. These results show two things: 1) the mutation rate in ZIKV was initially relatively low, probably less than \( 0.2 \times 10^{-3} / \text{site/year} \); but it increased significantly in many descendant lineages and 2) ZIKV appeared roughly in 1930–1945.

Accumulation of New Mutations in Asian Strains

Using the Phylogenetic Analyses by Maximum Likelihood (PAML) with Jones–Taylor–Thornton (JTT), Whelan and Goldman (WAG), and Dayhoff models of amino acid substitution (Yang 2007), amino acid changes at all branches of the PE tree (fig. 1B) were identified. The results consistently show that the ancestors of Aae-Malaysia, Philippine, Cambodia,
Thailand, Polynesia, and 24 American strains have accumulated 33, 14, 3, 5, 2, and 1 new amino acid changes, respectively (supplementary table S5, Supplementary Material online). In this inference, mutations that are polymorphic among Asian strains or those found in the Spondweni virus are excluded because it is possible that the mutant forms are actually ancestral forms. None of these mutations identified here was tested in the previous site-directed mutagenesis experiments using different flaviviruses (Kawano et al. 1993; Amberg and Rice 1999; Mandl et al. 2000; Crill et al. 2007; Zhou et al. 2007; Dong et al. 2010; Winkelmann et al. 2011; Zou et al. 2011; Huang et al. 2014) and some of these mutations could be responsible for causing serious medical problems in human populations. Recent physicochemical analyses suggest that these mutations have a low chance of causing medical problems (Faria et al. 2016). However, we have little knowledge about the individual and nonadditive effects of these mutations of ZIKV function, which need to be determined by mutagenesis experiments.

**RNA Structures**

Within the 5′-UTR, six mutations occurred in Aae-Malaysia and have been maintained virtually in all Asian strains (supplementary table S6, Supplementary Material online). The RNA-folding structures of the stem-loop (SL-A) of Aae-Malaysia, predicted by mfold modeling (Zuker 2003) are similar to that of MR766 (supplementary fig. S5A and B, Supplementary Material online) and no critical mutation can be identified. The 3′-UTR of ZIKV folds into five major SL structures (SL-I, SL-II, SL-III, SL-DB, and 3′-SL) (Zhu et al. 2016). In this region, 16 mutations can be found, 13 of which occurred in Aae-Malaysia and the other three in Philippine strains have been maintained in all of their descendants (supplementary table S6, Supplementary Material online). The predicted RNA structures of Aae-Malaysia are either identical (SL-II) or very similar (SL-III) to the structure of MR766 (supplementary fig. S5C and D, Supplementary Material online, respectively) and the mutations in these regions do not seem to modify the function of the Asian strains.

Mutations in the rest of the 3′-UTR have distinct possibilities of changing ZIKV functions. The RNA structure of the SL-I of MR766 consists of 24 nucleotides (positions 35 through 58, ΔG = −12.9 Kcal/mol, fig. 3A). Compared with this, Aae-Malaysia with a single mutation (u42c) and Philippine with an additional mutation (u9c) both generate more stable RNA structures with 24-nucleotides (ΔG = −15.3 Kcal/mol) or those with 58-nucleotides (ΔG = −14.8 to −15.6 Kcal/mol) (common segments in blue strips, fig. 3A). Functional importance of this region for flaviviral replication is indicated by chimeric mutants of Dengue type 1 virus (DENV1) (Tajima 2000). Functional importance of this region for flaviviral replication is indicated by chimeric mutants of Dengue type 1 virus (DENV1) (Tajima 2000).

### Table 1. Divergence Times and the Origins of Various ZIKV Strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Time (in years)</th>
<th>Evolutionary Rate (× 10^-3/site/year)</th>
<th>Strain of origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippines</td>
<td>16.2</td>
<td>a3 = 0.68</td>
<td>Philippine and Micronesia (1963)</td>
</tr>
<tr>
<td>Cambodia</td>
<td>32.0</td>
<td>a4 = 0.58</td>
<td>Cambodia (1995)</td>
</tr>
<tr>
<td>Thailand</td>
<td>4.3</td>
<td>a6 = 0.67</td>
<td>Thailand (1999)</td>
</tr>
<tr>
<td>Polynesia</td>
<td>7.9</td>
<td>a6 = 0.53</td>
<td>Polynesia (2007)</td>
</tr>
<tr>
<td>Brazil-Haiti-I and Suriname-I</td>
<td>6.2</td>
<td>a7 = 1.46</td>
<td>Brazil-I, Haiti-I and Suriname-I (2013)</td>
</tr>
<tr>
<td>Alu1</td>
<td>9.9</td>
<td>a13 = 3.79</td>
<td>Alu1 (1957)</td>
</tr>
<tr>
<td>Nigeria, Alu2 and Aaf4</td>
<td>8.0</td>
<td>a14 = 4.39</td>
<td>Nigeria, Alu2 and Aaf4 (1965)</td>
</tr>
<tr>
<td>Aaf1</td>
<td>3.7</td>
<td>a17 = 0.92</td>
<td>Aaf1 (1962)</td>
</tr>
<tr>
<td>Aaf2</td>
<td>11.6</td>
<td>a18 = 2.18</td>
<td>Aaf2 and Aaf3 (1962)</td>
</tr>
</tbody>
</table>

*10,176 comparable nucleotide sites were considered.

### Table 2. Evolutionary Rates of Nucleotide and Amino Acid Substitutions (× 10^-3/site/year).

<table>
<thead>
<tr>
<th>ZIKV</th>
<th>Divergence Time (years)</th>
<th>Position Within a Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First</td>
<td>Second</td>
</tr>
<tr>
<td>1. Aae-Malaysia</td>
<td>19.25</td>
<td>1.4 ± 0.14</td>
<td>0.3 ± 0.07**</td>
</tr>
<tr>
<td>2. Philippine</td>
<td>16.0</td>
<td>0.3 ± 0.07</td>
<td>0.1 ± 0.04**</td>
</tr>
<tr>
<td>3. Micronesia</td>
<td>44.0</td>
<td>0.3 ± 0.04</td>
<td>0.1 ± 0.03**</td>
</tr>
<tr>
<td>4. Cambodia</td>
<td>15.1</td>
<td>0.3 ± 0.07</td>
<td>0.1 ± 0.03**</td>
</tr>
<tr>
<td>5. Thailand</td>
<td>14.8</td>
<td>0.3 ± 0.08</td>
<td>0.1 ± 0.05**</td>
</tr>
<tr>
<td>6. Polynesia</td>
<td>6.2</td>
<td>0.2 ± 0.09</td>
<td>0.05 ± 0.05**</td>
</tr>
<tr>
<td>7. Brazil-I</td>
<td>1.3</td>
<td>0.9 ± 0.46</td>
<td>0.07</td>
</tr>
<tr>
<td>8. Haiti-I</td>
<td>1.1</td>
<td>0.8 ± 0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>9. Suriname-I</td>
<td>1.9</td>
<td>0.6 ± 0.32</td>
<td>0.2 ± 0.16**</td>
</tr>
<tr>
<td>10. Ada1</td>
<td>54</td>
<td>0.2 ± 0.03</td>
<td>0.2 ± 0.03**</td>
</tr>
<tr>
<td>11. Ada2</td>
<td>54</td>
<td>0.02 ± 0.01</td>
<td>0.01 ± 0.01**</td>
</tr>
<tr>
<td>12. Aaf1</td>
<td>21</td>
<td>0.6 ± 0.09</td>
<td>0.2 ± 0.05**</td>
</tr>
<tr>
<td>13. Alu1</td>
<td>32.1</td>
<td>1.3 ± 0.19</td>
<td>0.3 ± 0.08**</td>
</tr>
<tr>
<td>14. Nigeria</td>
<td>3.1</td>
<td>1.6 ± 0.36</td>
<td>1.3 ± 0.33**</td>
</tr>
<tr>
<td>15. Alu2</td>
<td>32.1</td>
<td>0.4 ± 0.06</td>
<td>0.1 ± 0.03**</td>
</tr>
<tr>
<td>16. Aaf4</td>
<td>19.1</td>
<td>0.1 ± 0.04</td>
<td>0.0 ± 0.02**</td>
</tr>
<tr>
<td>17. Aop</td>
<td>19.4</td>
<td>0.3 ± 0.07</td>
<td>0.0 ± 0.01**</td>
</tr>
<tr>
<td>18. Aaf2</td>
<td>3.8</td>
<td>0.8 ± 0.25</td>
<td>0.2 ± 0.11**</td>
</tr>
<tr>
<td>19. Aaf3</td>
<td>6.8</td>
<td>0.5 ± 0.14</td>
<td>0.1 ± 0.06**</td>
</tr>
</tbody>
</table>

*The evolutionary rate at the second position is smaller than that at the first position (P < 0.05).
**The evolutionary rate at the second position is smaller than those at the first and third positions (P < 0.01).
(*) and (**) The evolutionary rate at the second position is smaller than that at the third position (P < 0.05 and P < 0.01, respectively).
These observations strongly suggest that the seven critical mutations (u42c, c257u, u258c, and a266g, u424g, u426c, and c427u), which occurred in Aae-Malaysia prior to 1963, may be causing the currently known medical problems. It seems paradoxical that the association between ZIKV infection and microcephaly and other serious brain disorders has not been found prior to the 2013–2014 French Polynesian epidemic (Malone et al. 2016). However, it can be easily imagined that the ZIKV-disease associations were simply missed in earlier days, as we witnessed in the original survey of the French Polynesian epidemic. The geographic regions of ZIKV outbreaks coincide with the areas at risk of DENV and Chikungunya virus (CHIKV) outbreaks (Cao-Lormeau and Musso 2014; Shen et al. 2016). These viruses may cause various medical symptoms by interacting with each other (Fagbami et al. 1988; Géardin et al. 2014). Before, we can fully study such interactions in disease developments, it seems critical to clarify the role of each virus in causing a specific disease, like microcephaly. Fortunately, some animal models for ZIKV infection have been developed (Cugola et al. 2016; Li et al. 2016; Miner et al. 2016). At this stage of ZIKV research, the seven point mutations in the 3′-UTR, possibly together with some of the mutations that occurred in the polyproteins of Asian strains, offer a set of specific testable hypotheses.

**Fig. 3.** *mfold* RNA structure models of MR 766 and Asian strains. (A) The SL-I structures. (B) The SL-DB structures. (C) The 3′-SL structures. The comparable segments of the African and Asian strains are shown by blue strips, whereas the mutated sites are highlighted by yellow rectangles. The SL-I, SL-DB, and 3′-SL structures were modeled considering nucleotide sites 1–60, 237–322, and 332–428, respectively, and minor loops for the latter two are not shown.
(rather than using chimeric mutants) to examine how ZIKV started causing a variety of problems in brain development.

**Materials and Methods**

**Construction of Phylogenetic Trees of ZIKV Strains**

The “conventional” rooted evolutionary trees of 20 representative strains were constructed by applying ML, neighbour-joining (NJ), and maximum parsimony (MP) methods directly to their nucleotide sequence data together with the out-group, Spondweni virus (SM-6V-1), using the MEGA6 web server (Tamura et al. 2013). In constructing the ML and NJ trees, branch lengths ($d$) were estimated from the proportion of different nucleotides per site ($p_i$) using the relationship $d = - (3/4) \ln(1 - (4/3) p_i)$ under the JC model (Jukes and Cantor 1969) and the gamma distribution with the JC model; the standard error of $d$ was calculated from $\{[9p_i(1 - p_i)]/(3 - 4p_i)^2 n\}^{1/2}$, where $n$ is the number of nucleotide sites compared. The number of amino acid replacements per site ($k$) was estimated from the proportion of different amino acids per site ($p_j$), $K = \ln(1 - p_j)$, and the standard error was calculated from $[p_j/(1 - p_j)n]^{1/2}$, where $n$ is the number of amino acid compared (Nei and Kumar 2000). The Bayesian Evolutionary Analyses Sampling Trees (BEAST) analyses were also used to incorporate the variable times of virus sampling (or “time-stamped data”) (Drummond and Rambaut 2007). For each run, we tried to ensure and achieve the ESS of more than 200.

The new phylogenetic trees were obtained first by establishing the tree topology based on the shared amino acid replacements in their polyproteins (supplementary figs. S2 and S3, Supplementary Material online) and then by evaluating the numbers of nucleotide substitutions per site for all branches applying the ML method (Yang 2007). Using the ML-based Bayesian method (Yang 2007) with JTT, WAG, and Dayhoff models of amino acid substitution, ancestral amino acids at all nodes of the PE tree (fig. 2A) were inferred and amino acid replacements have been identified.

**Estimation of Divergence Times and the Evolutionary Rates**

The evolutionary rate of nucleotide substitution of each ZIKV strain was estimated by dividing its branch length by the time span that the virus was allowed to evolve. Divergence time estimates were based on the ML tree determined from the numbers of nucleotide substitution per codon (supplementary fig. S4B, Supplementary Material online). It was assumed that the ancestors of Ada1, Ada2, Aaf1, and Malaysia diverged from each other in April 1947. The divergence times ($T_{a_i}$) and evolutionary rates ($a_i$) (fig. 2A) were estimated in the following fashion.

For Asian strains, the branch length for Aae-Malaysia ($L_1$) was first evaluated from $(d_{10,1} - d_{10,12} + d_{11,1})/2$, where $d_{10,1}$, $d_{10,12}$, and $d_{11,1}$ are the branch lengths between MR766 (strain 0) and Aae-Malaysia (strain 1), between MR766 and Aaf1 (strain 12) and between Aae-Malaysia and Adf1, respectively (fig. 2B). Aae-Malaysia appeared in April 1947 and sampled in July 1966, or the divergence time of 19.25 years, and therefore $a_1$ is given by $L_1/19.25$ (supplementary table S2, Supplementary Material online). $L_1$ can also be evaluated using other formulae such as $(d_{10,1} + d_{11,1} - d_{10,12})/2$ and $(d_{11,1} + d_{11,12} - d_{11,12})/2$. However, some values of $T_{a_i}$ to $T_{a_9}$ become negative when the $L_1$ values are estimated using these formulae and thus they were not used.

To evaluate the evolutionary rates of Phillipine, Micronesia, and Cambodia, we need to estimate $T_1$ (fig. 2A). The branch lengths between node O and Micronesia ($L_3$) and between node O and Aae-Malaysia ($L_1$) were estimated from $(d_{13} - d_{11,12} + d_{11,1})/2$ and $(d_{13} + d_{11,12} - d_{11,12})/2$, respectively (fig. S2B). In this calculation, Micronesia, rather than Philippine or Cambodia, was used because Micronesia was sampled earlier than the other two strains, which minimize the possible effect of random genetic drift on the analyses. Then, using relationships $a_1 (19.25 - T_1) = L_1, a_3 (60.17 - T_1) = L_3$, we obtain $T_1 = 19.25 (L_1 - L_1)/L_1$ and $a_3 = L_3 (60.17 - T_1)$ (supplementary table S2, Supplementary Material online). Similarly, pairs of $T_2$ and $a_2$, $T_3$ and $a_3$, and $T_4$ and $a_4$ were evaluated together. $T_2$ can be evaluated using the branch length of Philippine or Cambodia; however, the branch leading to Cambodia was used to assure positive values for $T_9$, $T_{a_9}$, and $T_5$. Knowing these divergence times, $a_2$, $a_9$, and $a_9$ can also be evaluated (supplementary table S2, Supplementary Material online).

For African strains, $a_{10}$ (Ada1) and $a_{11}$ (Ada2) were estimated from $[(d_{10,11} + d_{10,12} - d_{11,12})/2]/54$ and $[(d_{10,1} + d_{10,12} + d_{11,12})/2]/54$, respectively (supplementary table S2, Supplementary Material online). For African group 2 and 3 strains (fig. 2B), $a_{12} = [((-d_{10,11} + d_{10,12} + d_{11,12})/2)/21]$ was first evaluated and then, using relationships $a_{12} (21 - T_{a_{12}}) = L_{12,1}$ and $a_{13} (21 - T_{a_{13}}) = L_{13,6}$ and $a_{13}$ were evaluated from $21(L_{13} - L_{12,1})/L_{13}$ and $21(L_{13} - T_{a_{13}})$, respectively, where $L_{13} = (-d_{11,12} + d_{11,13} + d_{12,13})/2$ and $L_{12,1} = (d_{11,12} - d_{11,13} + d_{12,13})/2$. Continuing the process, $T_7$ and $a_{16}$, $T_{a_9}$, and $a_{17}$ and finally $T_9$ and $a_{19}$ were evaluated together and, eventually, $a_{15}$, $a_{16}$, and $a_{19}$ were determined. In evaluating all of these evolutionary rates, the branch lengths at all positions of each codon were used so that the effects of sampling biases are minimized.

**Inferring Amino Acid and Nucleotide Substitutions and RNA-Folding Modeling**

Using the PAML with JTT, WAG, and Dayhoff substitution models (Yang 2007), amino acid replacements at all nodes of the PE tree (fig. 2A) were inferred initially by considering the 21 strains (indicated by *, supplementary table S1, Supplementary Material online) and the maintenance of these mutations in their descendants were studied considering all of the remaining 36 strains. Because of the restriction on the currently available data sequence, the patterns of mutation accumulation in the 5'- and 3'-UTR in Asian strains were constrained to 20 and 13 strains, respectively (indicated by † and ‡, respectively, supplementary table S1, Supplementary Material online) and the maintenance of these mutations in their descendants were studied considering all of the remaining 36 strains. Because of the restriction on the currently available sequence data, the patterns of mutation accumulation in the 5'- and 3'-UTR in Asian strains were constrained to 20 and 13 strains, respectively (indicated by † and ‡, respectively, supplementary table S1, Supplementary Material online).

RNA-folding structures of a stem-loop (SL-A) in the 5'-UTR and those of (SL-I, SL-II, SL-III, SL-DB, and 3'-SL) in the 3'-UTR were evaluated together. In modeling SL-I, SL-II, SL-III, and RNA-Folding Modeling using other formulae such as $(d_{10,1} + d_{11,1} - d_{10,12})/2$ and $(d_{11,1} + d_{11,12} - d_{11,12})/2$, however, some values of $T_{a_9}$ to $T_{a_9}$ become negative when the $L_1$ values are estimated using these formulae and thus they were not used.
III, SL-DB, and 3'-SL, nucleotide sites 1–106 for the 5'-UTR and 1–60, 101–170, 180–235, 237–322, and 332–428 are considered, respectively.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

We thank Ruth Yokoyama and two anonymous reviewers for their helpful comments as well as Michael Derry and Darin Melton at Emory University for resolving IT issues. This work was supported partially by National Institutes of Health (EY016400) and Emory University (S.Y.).

References


