2010

Evolution of the Sequence Composition of Flaviviruses

Alyxandria M. Schubert

Catherine Putonti
Loyola University Chicago, cputonti@luc.edu

Recommended Citation

Schubert, A and C Putonti. "Evolution of the Sequence Composition of Flaviviruses." Infections, Genetics, and Evolution 10(1), 2010.

Evolution of the Sequence Composition of *Flaviviruses*

Alyxandria M. Schubert[1] and Catherine Putonti[1,2,3*]

1 Department of Bioinformatics, Loyola University Chicago, Chicago, IL USA

2 Department of Biology, Loyola University Chicago, Chicago, IL USA

3 Department of Computer Science, Loyola University Chicago, Chicago, IL USA

* To whom correspondence should be addressed.

Email: cputonti@luc.edu

Fax number: 773-508-3646

Address: 1032 W. Sheridan Rd., Chicago, IL, 60660

Schubert, AM and Putonti, C. Evolution of the Sequence Composition of *Flaviviruses*. Infection, Genetics and Evolution.

Abstract

The adaption of pathogens to their host(s) is a major factor in the emergence of infectious disease and the persistent survival of many of the infectious diseases within the population. Since many of the smaller viral pathogens are entirely dependent upon host machinery, it has been postulated that they are under selection for a composition similar to that of their host. Analyses of sequence composition have been conducted for numerous small viral species including the *Flavivirus* genus. Examination of the species within this particular genus that infect vertebrate hosts revealed that sequence composition proclivities do not correspond with vector transmission as the evolutionary history of this species suggests. Recent sequencing efforts have generated complete genomes for many viral species including members of the *Flavivirus* genus. A thorough comparison of the sequence composition was conducted for all of the available *Flaviviruses* for which the complete genome is publicly available. This effort expands the work of previous studies to include new vector-borne species as well as members of the insect-specific group which previously have not been explored. Metrics, including mono-, di-, and trinucleotide abundances as well as N_C values and codon usage preferences, were explored both for the entire polyprotein sequence as well as for each individual coding region. Preferences for compositions correspond to host-range rather than evolutionary history; species which infect vertebrate hosts exhibited particular preferences similar to each other as well as in correspondence with their host's preferences. *Flaviviruses* which do not infect vertebrate hosts, however, did not show these proclivities, with the exception of the Kamiti River virus suggesting its recent (either past or present) infectivity of an unknown vertebrate host.

Keywords: RNA viruses; Flaviviruses; Sequence composition; Codon usage

INTRODUCTION

Given the compact nature of the genomes of many viral pathogens, acquisition of host-specific coding regions cannot be supported. Since many of these viral pathogens are entirely dependent upon host machinery, it has been postulated that they are under selection for a composition similar to that of their host. Host species exhibit compositional preferences as a consequence of many factors including: nucleotide abundances, DNA stacking energies, methylation, modification, replication, and repair mechanisms (Karlin, 1998; Xia and Yuen, 2005). Examination of mono-, di-, tri- and tetranucleotide usage within small viral genomes has revealed correspondence between the pathogen and host species. For example, it has been observed that CpG dinucleotides are under-represented for the majority of these species, regardless of the nature of their genome (Karlin *et al*., 1994; Rima and McFerran, 1997; Auewarakul, 2004; Shackelton *et al*., 2006; Sewatanon *et al*., 2007; Greenbaum *et al*., 2008; Tao *et al*., 2009). Moreover, the correspondence between the codon usage preferences of eukaryotic viruses and their host species has been the focus of many studies including a wide variety of DNA-based viruses, RNA-based viruses and retro-transcribing viruses (e.g. Karlin *et al*., 1990; Levin and Whittome, 2000; Jenkins and Holmes, 2003; Zhao *et al*., 2003; Adams *et al*., 2004; Gu *et al*., 2004; Zhou *et al*., 2005; Shackelton *et al*., 2006; Tsai *et al*., 2007; van Hemert *et al*., 2007; Jiang *et al*., 2008; Tao *et al*., 2009). Studies have supported both translational and mutational selection as the primary force shaping codon bias in these viral species (Jenkins *et al*., 2003). The correspondence of longer subsequences, four (Pride *et al*., 2006) or 17 to 26 nucleotide long sequences (Kerr and Boschetti, 2006), within pathogen and host genomes has also been considered, but with mixed conclusions.

The *Flavivirus* genus, one of the three genera within the *Flaviviridae* family, consists of over 70 different known species. These single-stranded positive sense RNA viruses (~ 11,000 bases) encoding for three structural – capsid (C), membrane (prM/M), envelope (E) – and seven non-structural proteins – NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5 – are spread world-wide, some regions having several different species coexisting within a single area. *Flaviviruses* are most often transmitted by arthropods (primarily mosquitoes and ticks) to a variety of vertebrate hosts. Moreover, the *flavivirus* genus also includes species with no known vector (NKV) as well as species which do not infect vertebrates, henceforth referred to as the insect-specific group. Phylogenies derived from sequence alignments of individual coding regions and the whole genomic sequences of available *flavivirus* species suggest that species within one vector group are typically more closely related than species of the other vector groups (Billoir, 2000; Gaunt *et al.*, 2001; Cook and Holmes, 2006). Analysis of the base composition and codon usage of the *flavivirus* genus has been previously conducted based upon partial NS5 gene sequences and the complete polyprotein sequence of 13 species (Jenkins *et al.*, 2001). This study found relationships between base compositions and vector specificity which the authors attribute to differences in mutational biases among *flaviviruses* (Jenkins *et al.*, 2001). Furthermore, this study revealed that *flaviviruses* exhibit dinucleotide and codon biases, but these biases do not covary with arthropod association (Jenkins *et al.*, 2001).

Since the study conducted by Jenkins *et al.* (2001), additional isolates have been sequenced as well as new species totaling well over 2,000 complete *flavivirus* sequences. Moreover, new species within the insect-specific group have been discovered and their complete genomes have been sequenced and annotated, including Cell Fusing Agent virus (Cammisa-Parks *et al.*, 1992), Kamiti River virus (Crabtree *et al.*, 2003), Culex flavivirus (Hoshino *et al.*,

2007) and Quang Binh virus (Crabtree *et al.*, 2009). This prompted us to revisit the study of Jenkins *et al*. (2001) and expand analysis to all of the publicly available annotated *flavivirus* species. Furthermore, because the individual protein products produced from the transcribed polyprotein vary in the number required to form the mature virion as well as their interaction with the host, it is likely that the pressures of selection for a composition similar to that of their host varies from gene to gene. Thus, in addition to examining the compositional properties of the NS5 gene, the compositional properties of each of the 10 *flavivirus* genes were examined.

MATERIALS AND METHODS

Flavivirus Genome Sequences. A total of 37 *flavivirus* genomes were used in this study (Table 1). These sequences were obtained from NCBI's RefSeq collection (http://www.ncbi.nlm.nih.gov/). The virus name and abbreviation, isolate, vector group, length (base pairs), and GenBank accession numbers are listed in Table 1. Protein coding regions were manually annotated according to their GenBank file annotations.

[Table 1]

Compositional Bias Measures. The mononucleotide usage of each of the genomes was calculated considering the overall GC content as well as the GC_1, GC_2, GC_3 and GC_{12} contents of the polyprotein coding regions, referring to the GC-content at first, second, third and first and second position of the codon respectively. The dinucleotide frequencies were also calculated. The dinucleotide relative abundances ρ^*_{XY} were computed according to the mononucleotide and dinucleotide frequencies as described in the study of Karlin and Burge (1995). The dinucleotide relative abundance value evaluates the differences between the observed dinucleotide frequencies and the expected frequencies determined by the constituent mononucleotide frequencies. Using this test statistic, $\rho^*_{XY} \leq 0.78$ characterizes extreme under-representation and $\rho^*_{XY} \geq 1.23$ characterizes extreme over-representation (Karlin and Burge, 1996). Calculations for the trinucleotide relative abundances were computed based upon the underlying dinucleotide (and subsequently underlying mononucleotide) abundances; for the trinucleotide *xyz*, the relative abundance is equal to $f_{xyz}/(f_{xy} \times f_{yz})$.

Codon Usage Measures. The "Effective Number of Codons" (ENC) was computed over the entire coding region of each *flavivirus* (Wright, 1990). The values of this test statistic, N_C, range from 20 (meaning only one codon is used for each amino acid) up to 61 (meaning all codons are used equally). The three stop codons are excluded from consideration. Additionally, the codon usage frequencies were calculated for each gene, including the three structural genes and the seven non-structural genes, in the 24 genomes for which the individual gene annotations were available (denoted with an asterisk in Table 1). These 24 sequences vary across vector groups (insect-specific, mosquito, tick and no known vector groups), providing a sufficient sample size. Calculations were also performed for the overall polyprotein for all 37 genomes. These biases were computed as f_{xyz} / f_{aa}, where f_{xyz} is the frequency of codon *xyz*, and f_{aa} is the frequency of the occurrence of all codons encoding for the amino acid of *xyz*. Thus for each species sequence (be it the complete polyprotein or an individual gene), a vector of the 61 biases was generated (excluding the three stop codons). In order to assess the similarity in the usage of particular codons, we calculated the Pearson correlation (*r*) for each pair of vectors generating a similarity matrix based solely on preferences in codon usage. Calculations were performed to compare these codon biases of a single gene between every pair combination in the 24 *flaviviruses* and likewise for each of the 37 individual polyprotein sequences. All of these calculations were conducted using code developed within our laboratory in C++.

Phylogeny Construction. To visualize the similarities in codon usage between the *flaviviruses*, phylogenies were derived based upon the correlations in their usages. The pair-wise distances between the individual genomes was calculated as (1-*r*)/2 such that anticorrelation (*r* = -1) is indicated with a distance of 1 and perfectly correlated genomes (*r* = 1) produce a distance of 0. Phylogenetic trees were generated using the FITCH application of the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) and visualized using PhyloWidget (Jordan and Piel, 2008). Because of its highly divergent sequence, the CFAV sequences were used as an outgroup in each phylogeny, in a manner similar to other phylogenies derived for the genus (e.g. Cook and Holmes, 2006). MODV and WNV-1 were excluded from the prM/M calculations due to inconclusive details within their annotations.

RESULTS

Compositional biases

The mono- and dinucleotide content for each of the 37 RefSeq genomic sequences listed in Table 1 was calculated. As previously observed with respect to GC content (Jenkins *et al.*, 2001), the GC content of coding regions varies according to the vector species. The lowest overall GC content is found in *flaviviruses* with no known vector (0.4384 ± 0.0010), mosquito and insect-specific groups have an intermediate GC content (0.4947 ± 0.0003 and 0.5162 ± 0.0004 respectively), and lastly the tick group has the highest GC content (0.5407 ± 0.0001). The mean overall GC content for all of *flaviviruses* considered in this study is 0.4981 ± 0.0004.

The ρ^*_{XY} value for the CpG dinucleotide is categorized as under-represented in 34 of the 37 *flaviviruses* (Table 2). This pattern of under-representation, however, is not present for QBV, CFAV and CxFV. Not surprisingly, these three *flaviviruses* are all classified within the insect-specific group and are not known to infect a vertebrate host. The KRV RefSeq genomic sequence does not fall below the threshold value (see Methods) exhibiting a minor suppression the CpG dinucleotide. In contrast to CpG, the TpA dinucleotide is found to be under-represented unanimously among all *flaviviruses* while the TpG dinucleotide is over-represented in all of *flaviviruses*. The ρ^*_{XY} values for some of the more interesting dinucleotides are summarized in Table 2.

[Table 2]

Codon Usage

Firstly, the GC-contents and N_C value was calculated for each of the 37 RefSeq polyproteins (see Methods). The results by species and by vector group are listed in Tables 3 and 4, respectively. As is seen in Table 4, the *flavivirus* group with the greatest N_C value, equating to the least bias in codon usage, is the insect-specific group (56.443 ± 0.050); mosquito and tick groups have an intermediate level of codon biases (52.397 ± 0.026 and 53.963 ± 0.016

respectively), and finally the NKV group has the highest level of codon biases (50.242 ± 0.075). Figure 1 shows the distribution of the GC_3 and N_C values according to the species' group. Because the effective codon usage for all of the species' polyprotein sequences was lower than the expectation, it can be concluded that base composition is not the only factor influencing codon usage. Again it can be seen that *flaviviruses* do not form distinct clusters by their respective vectors, but rather these groupings of various vectors are loosely organized and intersecting one another. For instance, although on average the NKV group displays low values of GC_3 and N_C, the Apoi virus (labeled A in Figure 1) displays GC_3 and N_C values similar to the mosquito group. The Dengue viruses (serotypes 1, 2, 3, and 4, indicated in Figure 1 as D1, D2, D3 and D4, respectively) exhibit GC_3 and N_C values similar to the NKV group. *Flaviviruses* with a tick vector overlap with those viruses with a mosquito vector. Lastly members of the insect-specific group fall closest along the expectation curve at various points.

[Figure 1]

[Table 3 & Table 4]

The availability of annotated genomes provides us with the opportunity to analyze not only the entire polyprotein of these annotated sequences but also each individual gene. The overall GC-content, GC_{12} and GC_3 values within each individual coding region are relatively constant between the *flavivirus* species. Each coding region exhibits an overall GC-content and GC_3 value around 50%. These individual statistics can be found in the supplemental data. The N_C

values for the NS5 coding region of the individual species was computed revealing a bias (52.26±2.78) comparable to that observed over the entire polyprotein.

The individual codon frequencies for both the entire polyprotein as well as each of the individual coding regions were also calculated. Phylogenetic trees based upon the correlation of codon preferences were derived (see Methods). In contrast with the more familiar alignment-based phylogenies, the codon-based phylogenies derived here provide a visualization of the relative similarity in codon usage amongst coding regions. For each pair of sequences being considered, the distance is calculated as the normalized correlation of codon usage (see Methods) such that two sequences exhibiting a similar codon usage preference have a smaller distance than two sequences preferring different codons.

The tree for the polyprotein region is shown in Figure 2. Based upon the codon usages within the polyprotein, species transmitted by the same vector do not form distinct clades in contrast with phylogenies derived from whole genome alignments (Cook and Holmes 2006). The hierarchy in Figure 2 appears to group species based upon the similarity in their overall GC-content. By reexamining the dinucleotide and trinucleotide compositional biases, however, it is evident that preferences (e.g. overrepresentation of TpG and CpA and under-representation of CpG and TpA) are consistent amongst the vertebrate host species. As such, the preferences associated with codons containing these dinucleotides are maintained across species. Thus, the GC-content cannot be the only factor influencing the observed similarity/dissimilarity in codon usage.

[Figure 2]

Similar to the analysis of the correspondence of the codon preferences within the polyprotein sequences, the correlations were calculated for each individual gene. From the tree of the NS5 gene (Figure 3), the most strongly conserved gene among the *flaviviruses* (Cook and Holmes, 2006), as well as the NS3 gene (Figure 4), it is once again apparent that the correlation in the codon frequencies does not correspond with the vector. In fact throughout each phylogeny found in this study there were no vector groups that formed well-defined clades. Trees for the other coding regions can be found in the supplementary data. In each case we see that although many times species exhibit a codon usage similar to that of other species in its same vector group, distinct clades for each individual vector group do not exist for any of the individual gene trees. This confirms that the codon usage does not covary with the vector group. Of particular interest, KRV displays interesting phylogenetic relationships with other *flaviviruses*. KRV, which was previously found to have the most similarity in terms of nucleotide sequence and growth kinetics in culture to CFAV (Cook and Holmes, 2006), consistently shows the most similarity with respect to codon usage patterns to *flaviviruses* capable of infecting vertebrates rather than with the other members of the insect-specific group. It is worth noting, that while the genome of KRV differs from the genomes of the other members insect-specific group in the length of its 3’ UTR region (Gritsun and Gould, 2006), this is not taken into consideration as only the coding regions are considered.

[Figure 3 & Figure 4]

DISCUSSION

The TpA dinucleotide is also unanimously under-represented throughout all of the *flaviviruses* in this study. This dinucleotide has been reported to be universally suppressed across all eukaryotes and prokaryotes with few exceptions (Ohno, 1988; Beutler *et al.,* 1989; Karlin and Burge, 1995). It has been proposed that the restriction of TpA is due to its low stacking energy (the lowest of all dinucleotides) or because of TpA's participation in many regulatory sequences (help avoid incorrect binding of regulatory factors) (Burge *et al.,* 1992).

The examination of the compositional properties of numerous vertebrate viruses has observed that CpG dinucleotides are under-represented for the majority of these species, regardless of the nature of their genome (Karlin *et al*., 1994; Rima and McFerran, 1997; Auewarakul, 2004; Shackelton *et al*., 2006; Sewatanon *et al*., 2007; Greenbaum *et al*., 2008; Tao *et al*., 2009). CpG suppression is common across all vertebrates, including mammals, and is commonly explained by the methylation-deamination-mutation mechanism which is described as the methylation of position 5 of the cytosine and the subsequent deamination of the newly formed 5-methylcytosine to produce thymine and conversion to TpG/CpA (Karlin and Burge, 1995). This suppression of CpG and related excess of TpG/CpA pattern is seen in the results listed in Table 2. This same tendency was observed within the *flavivirus* species which infect vertebrate hosts. Because *flaviviruses* cannot be methylated and yet most in this study do follow the methylation-deamination-mutation mechanism CpG and TpG/CpA levels as opposed to the pattern of other small viral genomes, it supports the hypothesis that these *flaviviruses* are adapting the dinucleotide usage patterns of its vertebrate host rather than invertebrate vector. Further confirming this hypothesis is the fact that this pattern of under-representation is not present for QBV, CxFV and CFAV which cannot infect mammalian cells (Grard *et al*., 2006;

Hoshino *et al*., 2007; Crabtree *et al*., 2009); CFAV, however is only slightly above the ρ*<0.78 threshold for under-representation determined by data simulations and statistical theory by Karlin and Burge (1996); Karlin *et al*. (1994), in fact, having conducted calculations using the same sequence, regard CFAV as under-representing CpG.

Intriguingly the $\rho*_{CG}$ value for the insect-specific virus KRV, which also cannot infect mammalian cells (Sang *et al*., 2003), is below the ρ*<0.78 threshold; its value of ρ*0.76 is a value similar to that of the human viruses hepatitis C and O'nyong-nyong virus (Karlin *et al*., 1994). The *flavivirus* which infect vertebrates have a significantly smaller $\rho*_{CG}$ value (on average 0.48±0.09). This suggests that perhaps (1) KRV used to be able to infect a vertebrate host and is under reduced selection to suppress CpG in its genome or (2) KRV is evolving towards the ability to infect a vertebrate host and is thus under selection to suppress CpG or (3) the value 0.78 for under-representation is not entirely an accurate threshold. In the case of the latter, this threshold was slightly relaxed to 0.8 when examining the complete genomes of 75 viruses infecting vertebrates with a genome size <30 Kbp; this study revealed that 71 of the viruses under-represented ($\rho*_{CG}$<0.8) CpG (Karlin *et al*., 1994). Subsequent studies (Greenbaum *et al*., 2008) using different metrics have found results analogous with those of Karlin *et al*. (1994). If *flaviviruses* are suppressing CpG solely because their host suppresses CpG (i.e. it does not aid the virus in its virulence), a trend that has been observed for a variety of pathogens and their host(s) (Greenbaum *et al*., 2008), then the first two hypotheses are equally plausible: either KRV used to (or occasionally does although it has not yet been observed) infect a vertebrate host or it is evolving towards infecting a vertebrate host which suppresses CpG, as has been observed over time for influenza A viruses which crossed from birds to humans (Greenbaum *et al*., 2008). Although the primary host species for all of the *flaviviruses* (Table 1) vary in their CpG suppression, a correlation between primary host species and viral species CpG suppression does not exist.

Another virus with unique dinucleotide composition is the TABV, which is exclusively over-representing the CpC and GpG dinucleotides. This is somewhat similar to the patterns found in Karlin and Burge (1995) in the excess of these dinucleotides within small viral genomes; however, it does not follow the normal range for the TpG/CpA dinucleotides, which in TABV are also over-represented. In the phylogenetic analysis of *flaviviruses* by Cook and Holmes (2006), TABV had to be omitted from phylogenies due to its highly genetically divergent structural genes.

The codon trees implemented here provide an alternative means of assessing the similarity between species taking into consideration mutational biases due to host range. Because definite vector groupings were not detected in any trees (Figures 2, 3, 4 and supplemental figures), the codon preferences observed amongst the 37 *flaviviruses* are not vector specific. The *flavivirus* group of each species does not determine codon preference, but rather the host type (vertebrate or invertebrate) plays the most influential role in determining codon biases. Moreover, similarities in composition within the mosquito-vector viral species are not distinguishable between species of the "*Culex* clade" (those transmitted by *Culex* mosquitoes) and the "*Aedes* clade" (those transmitted by *Aedes* mosquitoes), a result of traditional alignment-based phylogenies (e.g. Gaunt *et al*., 2001). The compositional biases within the viruses which have a primary vertebrate host species exhibit a more similar codon usage with other species with a primary vertebrate host than they do with the members of the insect-specific group. The exception is the insect-specific species KRV which exhibits a codon usage more closely related to the vertebrate host species. This, in addition to the observed usage of CpG within KRV, suggests that this species used to be or recently became under selection to adapt to a vertebrate composition.

While theories on the evolution of this genus include (1) divergence of the NKV species prior to vector transmission or (2) NKV species and tick-borne species as sister groups (Cook and Homes, 2006), neither trend is observed when considering codon usage preferences within species sharing the same vector. In contrast, trees based upon the NS5 and NS3 amino acid sequences support theories (1) and (2), respectively. By examining the amino acid sequences, however, more subtle forms of selection (such as selecting for the codon preference of the host) are not considered. If in fact the genus evolved following one of these two histories from an ancestor species unable to infect vertebrates, the compositional proclivities of the individual species' genomes are a result of selection based upon their host species. As the codon trees derived here show, species known to have evolved from a common ancestor (such as the dengue virus serotypes) do not always appear monophyletic. This suggests that the organization of the nodes within the codon trees is not representative of the evolution of the species as we see from conventional amino acid alignment-based trees, but rather of the selection upon the species to adapt to its particular host. Thus, expanding this work to include the genomes of additional *flavivirus* strains not belonging to the RefSeq collection would not likely result in clustering based upon species. The specific hosts as well as length of time infecting a particular host are significant variables that determine the extent to which the *flavivirus* has adopted its host's codon preferences.

REFERENCES

1. Adams, M.J., Antoniw, J.F., 2004. Codon usage bias among plant viruses. Arch. Virol. 149, 113-135.
2. Auewarakul, P., 2004. Composition bias and genome polarity of RNA viruses. Virus Res. 109, 33-37.
3. Baillie, G.J., Kolokotronis, S.O., Waltari, E., Maffei, J.G., Kramer, L.D., Perkins, S.L., 2008. Phylogenetic and evolutionary analyses of St. Louis encephalitis virus genomes. Mol. Phylogenet. Evol. 47, 717-728.
4. Bakonyi, T., Gould, E.A., Kolodziejek, J., Weissenböck, H., Nowotny, N., 2004. Complete genome analysis and molecular characterization of Usutu virus that emerged in Austria in 2001: comparison with the South African strain SAAR-1776 and other flaviviruses. Virology 328, 301-310.
5. Beutler, E., Gelbart, T., Han, J.H., Koziol, J.A., Beutler, B., 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. U.S.A. 86, 192-196.
6. Billoir, F., de Chesse, R., Tolou, H., Micco, P., Gould, E.A., de Lamballarie, X., 2000. Phylogeny of the genus *Flavivirus* using complete coding sequences of arthropod-borne viruses and viruses with no known vector. J. Gen. Virol. 81, 781–790.
7. Burge, C., Campbell, A.M., Karlin, S., 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. U.S.A. 89, 1358-1362.
8. Cammisa-Parks, H., Cisar, L.A., Kane, A., Stollar, V., 1992. The complete nucleotide sequence of cell fusing agent (CFA): homology between the nonstructural proteins

encoded by CFA and the nonstructural proteins encoded by arthropod-borne flaviviruses. Virology 189,511-524.

9. Campbell, M.S., Pletnev, A.G., 2000. Infectious cDNA clones of Langat tick-borne flavivirus that differ from their parent in peripheral neurovirulence. Virology 269, 225-237.
10. Charlier, N., Leyssen, P., Pleij, C.W., Lemey, P., Billoir, F., Van Laethem, K., Vandamme, A.M., De Clercq, E., de Lamballerie, X., Neyts, J., 2002. Complete genome sequence of Montana Myotis leukoencephalitis virus, phylogenetic analysis and comparative study of the 3' untranslated region of flaviviruses with no known vector. J. Gen. Virol. 83, 1875-1885.
11. Charrel, R.N., Zaki, A.M., Attoui, H., Fakeeh, M., Billoir, F., Yousef, A.I., de Chesse, R., De Micco, P., Gould, E.A., de Lamballerie, X., 2001. Complete coding sequence of the Alkhurma virus, a tick-borne flavivirus causing severe hemorrhagic fever in humans in Saudi Arabia. Biochem. Biophys. Res. Commun. 287, 455-461.
12. Coimbra, T.L., Nassar, E.S., Nagamori, A.H., Ferreira, I.B., Pereira, L.E., Rocco, I.M., Ueda-Ito, M., Romano, N.S., 1993. Iguape: a newly recognized flavivirus from São Paulo State, Brazil. Intervirology 36, 144-152.
13. Cook, S., Holmes, E.C., 2006. A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: *Flaviviridae*) and the evolution of vector transmission. Arch. Virol. 151, 309-325.
14. Crabtree, M.B., Nga, P.T., Miller, B.R., 2009. Isolation and characterization of a new mosquito flavivirus, Quang Binh virus, from Vietnam. Arch. Virol. 154, 857-860.

15. Crabtree, M.B., Sang, R.C., Stollar, V., Dunster, L.M., Miller, B.R., 2003. Genetic and phenotypic characterization of the newly described insect flavivirus, Kamiti River virus. Arch. Virol. 148, 1095-1118.

16. de Lamballerie, X., Crochu, S., Billoir, F., Neyts, J., de Micco, P., Holmes, E.C., Gould, E.A., 2002. Genome sequence analysis of Tamana bat virus and its relationship with the genus Flavivirus. J. Gen. Virol. 83, 2443-2454.

17. Durbin, A.P., Karron, R.A., Sun, W., Vaughn, D.W., Reynolds, M.J., Perreault, J.R., Thumar, B., Men, R., Lai, C.J., Elkins, W.R., Chanock, R.M., Murphy, B.R., Whitehead, S.S., 2001. Attenuation and immunogenicity in humans of a live dengue virus type-4 vaccine candidate with a 30 nucleotide deletion in its 3'-untranslated region. Am. J. Trop. Med. Hyg. 65, 405-413.

18. Gao, G.F., Hussain, M.H., Reid, H.W., Gould, E.A., 1994. Identification of naturally occurring monoclonal antibody escape variants of louping ill virus. J. Gen. Virol. 75, 609-614.

19. Gaunt, M.W., Sall, A.A., de Lamballarie, X., Falconar, A.K.I., Dzihivanian, T.I., Gould, E.A., 2001. Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. J. Gen. Virol. 82, 1867–1876.

20. Grard, G., Lemasson, J.J., Sylla, M., Dubot, A., Cook, S., Molez, J.F., Pourrut, X., Charrel, R., Gonzalez, J.P., Munderloh, U., Holmes, E.C., de Lamballerie, X., 2006. Ngoye virus: a novel evolutionary lineage within the genus *Flavivirus*. J. Gen. Virol. 87, 3273-3277.

21. Greenbaum, B.D., Levine, A.J., Bhanot, G., Rabadan, R., 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4, e1000079.

22. Gritsun, T.S., Gould, E.A., 2006. The 3' untranslated regions of Kamiti River virus and Cell fusing agent virus originated by self-duplication. J. Gen. Virol. 87, 2615-2619.

23. Gritsun, T.S., Venugopal, K., Zanotto, P.M., Mikhailov, M.V., Sall, A.A., Holmes, E.C., Polkinghorne, I., Frolova, T.V., Pogodina, V.V., Lashkevich, V.A., Gould, E.A., 1997. Complete sequence of two tick-borne flaviviruses isolated from Siberia and the UK: analysis and significance of the 5' and 3'-UTRs. Virus Res 49, 27-39.

24. Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155-161.

25. Hoshino, K., Isawa, H., Tsuda, Y., Yano, K., Sasaki, T., Yuda, M., Takasaki, T., Kobayashi, M., Sawabe, K., 2007. Genetic characterization of a new insect flavivirus isolated from *Culex pipiens* mosquito in Japan. Virology 359, 405-414.

26. Hurrelbrink, R.J., Nestorowicz, A., McMinn, P.C., 1999. Characterization of infectious Murray Valley encephalitis virus derived from a stably cloned genome-length cDNA. J. Gen. Virol. 80, 3115-3125.

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

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

29. Jiang, Y., Deng, F., Wang, H., Hu, Z., 2008. An extensive analysis on the global codon usage pattern of baculoviruses. Arch. Virol. 153, 2273-2282.

30. Jordan, G.E., Piel, W.H., 2008. PhyloWidget: web-based visualizations for the tree of life. Bioinformatics 24, 1641-1642.

31. Karlin, S., Blaisdell, B.E., Schachtel, G.A., 1990. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. J. Virol. 64, 4264-4273.

32. Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283-290.

33. Karlin, S., Doerfler, W., Cardon, L.R., 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. 68, 2889-2897.

34. Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1, 598-610.

35. Kerr, J.R., Boschetti, N., 2006. Short regions of sequence identity between the genomes of human and rodent parvoviruses and their respective hosts occur within host genes for the cytoskeleton, cell adhesion and Wnt signalling. J. Gen. Virol. 87, 3567-3575.

36. Kinney, R.M., Butrapet, S., Chang, G.J., Tsuchiya, K.R., Roehrig, J.T., Bhamarapravati, N., Gubler, D.J., 1997. Construction of infectious cDNA clones for dengue 2 virus: strain 16681 and its attenuated vaccine derivative, strain PDK-53. Virology 230, 300-308.

37. Kuno, G., Chang, G.J., 2006. Characterization of Sepik and Entebbe bat viruses closely related to yellow fever virus. Am. J. Trop. Med. Hyg. 75, 1165-1170.

38. Kuno, G., Chang, G.J., 2007. Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. Arch. Virol. 152, 687-696.

39. Laemmert, H.W., Hughes, T.P., 1947. The virus of Ilheus encephalitis: Isolation, serological specificity and transmission. J. Immunol. 55, 61-67.

40. Levin, D.B., Whittome, B., 2000. Codon usage in nucleopolyhedroviruses. J. Gen. Virol. 81, 2313-2325.

41. Leyssen, P., Charlier, N., Lemey, P., Billoir, F., Vandamme, A.M., De Clercq, E., de Lamballerie, X., Neyts, J., 2002. Complete genome sequence, taxonomic assignment, and comparative analysis of the untranslated regions of the Modoc virus, a flavivirus with no known vector. Virology 293, 125-140.

42. Lin, D., Li, L., Dick, D., Shope, R.E., Feldmann, H., Barrett, A.D., Holbrook, M.R., 2003. Analysis of the complete genome of the tick-borne flavivirus Omsk hemorrhagic fever virus. Virology 313, 81-90.

43. Mackenzie, J.S., Williams, D.T., 2009. The zoonotic flaviviruses of Southern, South-Eastern and Eastern Asia, and Australasia: The potential for emergent viruses. Zoonoses Public Health May 20 [Epub ahead of print]

44. Mandl, C.W., Heinz, F.X., Stöckl, E., Kunz, C., 1989. Genome sequence of tick-borne encephalitis virus (Western subtype) and comparative analysis of nonstructural proteins with other flaviviruses. Virology 173, 291-301.

45. Mandl, C.W., Holzmann, H., Kunz, C., Heinz, F.X., 1993. Complete genomic sequence of Powassan virus: evaluation of genetic elements in tick-borne versus mosquito-borne flaviviruses. Virology 194, 173-184.

46. Ohno, S., 1988. Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. Proc. Natl. Acad. Sci. U.S.A. 85, 9630-9634.

47. Peyrefitte, C.N., Couissinier-Paris, P., Mercier-Perennec, V., Bessaud, M., Martial, J., Kenane, N., Durand, J.P., Tolou, H.J., 2003. Genetic characterization of newly

reintroduced dengue virus type 3 in Martinique (French West Indies). J Clin Microbiol. 41, 5195-5198.

48. Pride, D.T., Wassenaar, T.M., Ghose, C., Blaser, M.J., 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8.
49. Puri, B., Nelson, W.M., Henchal, E.A., Hoke, C.H., Eckels, K.H., Dubois, D.R., Porter, K.R., Hayes, C.G., 1997. Molecular analysis of dengue virus attenuation after serial passage in primary dog kidney cells. J. Gen. Virol. 78, 2287-2291.
50. Rice, C.M., Lenches, E.M., Eddy, S.R., Shin, S.J., Sheets, R.L., Strauss, J.H., 1985. Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. Science 229, 726-733.
51. Rima, B.K., McFerran, N.V., 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J. Gen. Virol. 78, 2859-2870.
52. Sang, R.C., Gichogo, A., Gachoya, J., Dunster, M.D., Ofula, V., Hunt, A.R., Crabtree, M.B., Miller, B.R., Dunster, L.M., 2003. Isolation of a new flavivirus related to cell fusing agent virus (CFAV) from field-collected flood-water *Aedes* mosquitoes sampled from a dambo in central Kenya. Arch. Virol. 148, 1085-1093.
53. Sewatanon, J., Srichatrapimuk, S., Auewarakul, P., 2007. Compositional bias and size of genomes of human DNA viruses. Intervirology 50, 123-132.
54. Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551-563.
55. Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28-38.

56. Sumiyoshi, H., Mori, C., Fuke, I., Morita, K., Kuhara, S., Kondou, J., Kikuchi, Y., Nagamatu, H., Igarashi, A., 1987. Complete nucleotide sequence of the Japanese encephalitis virus genome RNA. Virology 161, 497-510.

57. Tajima, S., Takasaki, T., Matsuno, S., Nakayama, M., Kurane, I., 2005. Genetic characterization of Yokose virus, a flavivirus isolated from the bat in Japan. Virology 332, 38-44.

58. Tao, P., Dai, L., Luo, M., Tang, F., Tien, P., Pan, Z., 2009. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38, 104-112.

59. Theiler, M., Smith, H.H., 1937. The effect of prolonged cultivation in vitro upon the pathogenicity of yellow fever virus. J. Exp. Med. 65, 767-786.

60. Tsai, C.T., Lin, C.H., Chang, C.Y., 2007. Analysis of codon usage bias and base compositional constraints in iridovirus genomes. Virus Res. 126, 196-206.

61. Turell, M.J., Whitehouse, C.A., Butler, A., Baldwin, C., Hottel, H., Mores, C.N., 2008. Assay for and replication of Karshi (mammalian tick-borne flavivirus group) virus in mice. Am. J. Trop. Med. Hyg. 78, 344-347.

62. van Hemert, F.J., Berkhout, B., Lukashov, V.V., 2007. Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology 361, 447-454.

63. Weissenböck, H., Hubálek, Z., Bakonyi, T., Nowotny, N., 2009. Zoonotic mosquito-borne flaviviruses: Worldwide presence of agents with proven pathogenicity and potential candidates of future emerging diseases. Vet. Microbiol. August 26 [Epub ahead of print].

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

65. Xia, X., Yuen, K.Y., 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC Genet. 6, 20.

66. Yamshchikov, V.F., Wengler, G., Perelygin, A.A., Brinton, M.A., Compans, R.W., 2001. An infectious clone of the West Nile flavivirus. Virology 281, 294-304.

67. Zhao, K.N., Liu, W.J., Frazer, I.H., 2003. Codon usage bias and A+T content variation in human papillomavirus genomes. Virus Res. 98, 95-104.

68. Zhou, T., Gu, W., Ma, J., Sun, X., Lu, Z., 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. BioSystems 81, 77-86.

TABLES

Table 1. RefSeq viral genomes examined.

Virus	Isolate	Strain Origin	Principal Host Species	Vector Group	Length (bp)	Accession No.
Cell fusing agent (CFAV)*	-	*A. aegypti* [a]	*A. aegypti*	I	10695	NC_001564
Culex flavivirus (CxFV)	Tokyo	*C. pipiens* [a]	*C. pipiens*	I	10837	NC_008604
Kamiti River (KRV)*	SR-82	*A. macintoshi* [a]	*A. macintoshi*	I	11375	NC_005064
Quang Binh (QBV)	VN180	*C. tritaeniorhyncus* [a]	*C. tritaeniorhyncus*	I	10865	NC_012671
Bagaza (BAGV)	DakAr B209	mosquito [b]	Unknown	M	10941	NC_012534
Bussuquara (BSQV)	BeAn 4073	*A. belzebul* [U]	rodents, human	M	10290	NC_009026
Dengue virus type 1 (DENV-1)*	45AZ5	human [c,d]	human	M	10735	NC_001477
Dengue virus type 2 (DENV-2)*	16681	human [e,f,g,h,d,a]	human	M	10723	NC_001474
Dengue virus type 3 (DENV-3)*	D3/H/IMTSSA-SRI/2000/1266	human [i,a]	human	M	10707	NC_001475
Dengue virus type 4 (DENV-4)*	rDEN4	*A. pseudoscutellaris* [f,a,j]	human	M	10649	NC_002640
Entebbe bat (ENTV)	UgIL-30	*T. (C.) limbata* [b]	bat	M	10510	NC_008718
Iguape (IGUV)	SPAn 71686	sentinel mouse [U]	rodents, birds	M	10251	NC_009027
Ilheus (ILHV)*	Original	*Aedes* and *Psorophora* [U]	bird	M	10275	NC_009028
Japanese encephalitis (JEV)*	JaOArS982	mosquito [a]	bird, pig, human	M	10976	NC_001437
Kedougou (KEDV)	DakAar D1470	mosquito [b]	Unknown	M	10723	NC_012533
Kokobera (KOKV)	AusMRM 32	*C. annulirostris* [U]	kangaroos, horses	M	10233	NC_009029
Murray Valley encephalitis (MVEV)*	Australia 1951	human [a,j]	bird	M	11014	NC_000943
Sepik (SEPV)	MK7148	*M. septempunctata* [b]	Unknown	M	10793	NC_008719
St. Louis encephalitis (SLEV)	Kern217	*C. tarsalis* [j]	bat	M	10940	NC_007580
Usutu (USUV)*	Vienna 2001	blackbird [j]	bird	M	11066	NC_006551
West Nile virus lineage I (WNV-1)*	NY99	*B. scandiacus* [j]	bird	M	11029	NC_009942
West Nile virus lineage II (WNV-2)*	956	human [b,k,j,a]	bird	M	10962	NC_001563
Yellow fever (YFV)*	17D vaccine strain	human [l,m,j,k]	monkeys	M	10862	NC_002031
Yokose (YOKV)	Oita 36	bat [b,j]	bat	M	10857	NC_005039
Zika (ZIKV)	MR 766	sentinel monkey [b]	monkeys	M	10794	NC_012532
Apoi (APOIV)*	ApMAR	*Apodemus* [b,j]	rodents	NKV	10116	NC_003676
Modoc (MODV)*	M544	*P. maniculatus* [j]	bat	NKV	10600	NC_003635

Montana myotis leukoencephalitis (MMLV)*	Montana 1958	*M. lucifugus* ^{j}	bat	NKV	10690	NC_004119
Rio Bravo (RBV)*	RiMAR	bat ^{b,j}	bat	NKV	10140	NC_003675
Tamana bat (TABV)*	Tr127154	*P. parnellii* ^{n,j}	bat	NKV	10053	NC_003996
Alkhurma (AHFV)*	1176	human ^{b,j,o,n}	camels, sheep	T	10685	NC_004355
Karshi (KSIV)	LEIV 2247	*H. asiaticum asiaticum* ^{j}	rodents	T	10653	NC_006947
Langat (LGTV)*	TP21	*Ixodides* ^{j}	rodents	T	10943	NC_003690
Louping ill (LIV)*	369/T2	*I. ricinus* ^{b}	sheep	T	10871	NC_001809
Omsk hemorrhagic fever (OHFV)*	Bogoluvovska	*D. marginatus* ^{j}	muskrats	T	10787	NC_005062
Powassan (POWV)*	LB	human ^{b,p}	small mammals	T	10839	NC_003687
Tick-borne encephalitis (TBEV)*	Neudoerfl	tick ^{p}	rodents	T	11141	NC_001672

I indicates those species which do not infect a vertebrate host as the insect-specific group, M indicates those transmitted by a mosquito vector, T indicates those transmitted by a tick vector, NKV indicates those with no known vector. Asterisk denotes genomes for which individual gene annotations were available. Passage history of isolate includes: ^{a} C6/36 cells, ^{b} suckling mouse brain, ^{c} rhesus lung cells (FRhL), ^{d} PDK cells, ^{e} BS-C-1 cells, ^{f} LLC-MK$_2$, ^{g} *Rhesus macaque* monkey, ^{h} *T. amboinensis* mosquitoes, ^{i} white blood cells, ^{j} vero cells, ^{k} BHK-21 cells, l mouse embryonic tissue, m monkey serum in tyrode, ^{n} mouse, ^{o} sheep, ^{p} chick embryo, ^{U} unknown. Isolate and passage information and primary host information was ascertained from genome sequence publications (Baillie *et al*., 2008; Bakonyi *et al*., 2004; Billoir *et al*., 2000; Cammisa-Parks *et al*., 1992; Campbell & Pletnev, 2000; Charlier *et al*., 2002; Charrel *et al*., 2001; Coimbra *et al*., 1993; Crabtree *et al*., 2009; de Lamballerie *et al*., 2002; Durbin *et al*., 2001; Gao *et al*., 1994; Gritsun *et al*., 1997; Hoshino *et al*., 2007; Hurrelbrink *et al*., 1999; Kinney *et al*., 1997; Kuno and Chang, 2006; Kuno and Chang, 2007; Laemmert and Hughes, 1947; Leyssen *et al*., 2002; Lin *et al*., 2003; Mackenzie and Williams, 2009; Mandl *et al*., 1989; Mandl *et al*., 1993; Peyrefitte *et al*., 2003; Puri *et al*., 1997; Rice *et al*., 1985; Sang *et al*., 2003; Sumiyoshi *et al*, 1987; Tajima *et al*., 2005; Theiler and Smith, 1937; Turell *et al*., 2008; Weissenböck *et al*., 2009; Yamshchikov *et al*., 2001) and NCBI genome files.

Table 2. ρ^*_{XY} values for select dinucleotides that display over/under-representation.

Accession No.	Virus	Vector Group	ρ^*_{TA}	ρ^*_{TG}	ρ^*_{CA}	ρ^*_{CT}	ρ^*_{CC}	ρ^*_{CG}	ρ^*_{GC}	ρ^*_{GG}
NC_001564	Cell fusing agent virus	I	0.587	**1.232**	1.153	1.037	1.037	0.798	0.888	1.077
NC_008604	Culex flavivirus	I	0.495	**1.300**	1.194	0.975	0.917	0.931	0.910	0.927
NC_005064	Kamiti River virus	I	0.535	**1.310**	1.169	1.030	1.062	0.755	0.796	1.069
NC_012671	Quang Binh virus	I	0.532	**1.299**	1.158	1.003	0.953	0.905	0.913	0.968
NC_012534	Bagaza virus	M	0.548	**1.420**	**1.281**	1.069	1.118	0.570	0.901	1.042
NC_009026	Bussuquara virus	M	0.514	**1.382**	**1.251**	1.191	1.141	0.494	0.966	1.078
NC_001477	Dengue virus type 1	M	0.600	**1.403**	**1.260**	1.157	1.119	0.452	0.865	1.101
NC_001474	Dengue virus type 2	M	0.584	**1.392**	**1.277**	1.145	1.130	0.411	0.870	1.093
NC_001475	Dengue virus type 3	M	0.585	**1.397**	**1.305**	1.168	1.104	0.404	0.909	1.143
NC_002640	Dengue virus type 4	M	0.541	**1.369**	**1.288**	1.118	1.228	0.384	0.808	1.122
NC_008718	Entebbe bat virus	M	0.489	**1.488**	**1.263**	1.125	1.051	0.600	0.925	1.065
NC_009027	Iguape virus	M	0.450	**1.403**	**1.307**	1.153	1.037	0.540	0.848	1.099
NC_009028	Ilheus virus	M	0.424	**1.456**	**1.348**	1.134	1.023	0.545	0.919	1.097
NC_001437	Japanese encephalitis virus	M	0.557	**1.364**	1.223	1.217	1.049	0.584	0.957	1.029
NC_012533	Kedougou virus	M	0.498	**1.449**	**1.270**	**1.252**	1.121	0.496	0.979	1.021
NC_009029	Kokobera virus	M	0.531	**1.378**	1.165	1.169	1.216	0.555	0.928	1.023
NC_000943	Murray Valley encephalitis virus	M	0.543	**1.419**	**1.287**	1.092	1.143	0.514	0.995	1.011
NC_008719	Sepik virus	M	0.562	**1.467**	**1.282**	1.150	1.148	0.438	0.889	1.088
NC_007580	St. Louis encephalitis virus	M	0.457	**1.499**	**1.300**	1.108	1.093	0.533	0.929	1.032
NC_006551	Usutu virus	M	0.521	**1.376**	**1.284**	1.164	1.100	0.520	0.921	1.049
NC_009942	West Nile virus (lineage I)	M	0.483	**1.414**	**1.295**	1.159	1.034	0.574	0.918	0.991
NC_001563	West Nile virus (lineage II)	M	0.490	**1.438**	**1.239**	1.203	1.043	0.579	0.923	0.991
NC_002031	Yellow fever virus	M	0.449	**1.474**	**1.299**	**1.235**	1.187	0.383	0.882	1.085
NC_005039	Yokose virus	M	0.572	**1.447**	**1.234**	1.147	1.102	0.518	0.887	1.109
NC_012532	Zika virus	M	0.529	**1.435**	**1.304**	**1.255**	1.129	0.427	0.913	1.042
NC_003676	Apoi virus	NKV	0.505	**1.421**	**1.281**	**1.264**	1.104	0.411	0.913	1.042
NC_003635	Modoc virus	NKV	0.535	**1.469**	**1.301**	**1.260**	1.192	0.306	0.925	1.077
NC_004119	Montana myotis leukoencephalitis virus	NKV	0.525	**1.438**	**1.279**	**1.255**	1.172	0.305	0.873	1.107
NC_003675	Rio Bravo virus	NKV	0.519	**1.487**	**1.321**	1.201	1.070	0.349	0.888	1.072
NC_003996	Tamana bat virus	NKV	0.661	**1.330**	**1.255**	1.105	**1.315**	0.221	0.768	**1.249**

NC_004355	Alkhurma virus	T	0.488	**1.353**	**1.310**	**1.241**	1.083	0.538	0.873	1.062
NC_006947	Karshi virus	T	0.414	**1.445**	**1.347**	1.117	1.173	0.539	0.883	1.028
NC_003690	Langat virus	T	0.432	**1.451**	**1.360**	**1.262**	1.088	0.491	0.873	1.031
NC_001809	Louping ill virus	T	0.404	**1.427**	**1.313**	**1.290**	1.018	0.561	0.903	1.021
NC_005062	Omsk hemorrhagic fever virus	T	0.456	**1.441**	**1.281**	**1.287**	1.046	0.545	0.934	1.019
NC_003687	Powassan virus	T	0.445	**1.400**	**1.353**	1.183	1.105	0.515	0.904	1.061
NC_001672	Tick-borne encephalitis virus	T	0.417	**1.442**	**1.288**	**1.282**	1.049	0.551	0.903	1.041

Bold indicates over-represented dinucleotides; underline indicates under-represented dinucleotides.

Table 3. Viruses with their respective N_C, GC_3 levels, polyprotein length and vector.

Virus	N_C	GC_3	Polyprotein length (bp)	Vector
Cell fusing agent virus	57.54	0.571	9909	I
Culex flavivirus	54.44	0.603	10092	I
Kamiti River virus	57.89	0.541	10071	I
Quang Binh virus	55.90	0.586	10080	I
Bagaza virus	51.56	0.529	10281	M
Bussuquara virus	53.78	0.521	10290	M
Dengue virus type 1	49.87	0.462	10107	M
Dengue virus type 2	48.62	0.457	10104	M
Dengue virus type 3	48.92	0.468	10101	M
Dengue virus type 4	50.44	0.481	10092	M
Entebbe bat virus	54.30	0.567	10236	M
Iguape virus	50.56	0.557	10251	M
Ilheus virus	53.45	0.581	10188	M
Japanese encephalitis virus	54.99	0.556	10227	M
Kedougou virus	50.88	0.595	10227	M
Kokobera virus	53.14	0.527	10233	M
Murray Valley encephalitis virus	53.23	0.492	10233	M
Sepik virus	53.28	0.483	10218	M
St. Louis encephalitis virus	51.78	0.524	10293	M
Usutu virus	54.44	0.550	10233	M
West Nile virus (lineage I)	53.17	0.560	10299	M
West Nile virus (lineage II)	54.05	0.550	10221	M
Yellow fever virus	52.88	0.536	10164	M
Yokose virus	54.46	0.485	10278	M
Zika virus	52.54	0.541	10260	M
Apoi virus	54.34	0.501	10044	NKV
Modoc virus	49.73	0.447	10053	NKV
Montana myotis leukoencephalitis virus	49.84	0.415	10053	NKV
Rio Bravo virus	50.40	0.408	10068	NKV
Tamana bat virus	46.91	0.364	9975	NKV
Alkhurma virus	54.79	0.580	10179	T
Karshi virus	53.26	0.608	10251	T
Langat virus	53.36	0.592	10113	T
Louping ill virus	53.75	0.606	10176	T
Omsk hemorrhagic fever virus	53.38	0.576	10173	T
Powassan virus	54.86	0.573	10176	T
Tick-borne encephalitis virus	54.34	0.595	10173	T

Table 4. GC-contents and N_C values for the polyproteins of the different *flavivirus* groups.

Group	N_C	GC_3	GC_{12}	Overall GC
Insect-specific	56.443 (±0.050)	0.575	0.4866	0.5162
Mosquito	52.397 (±0.026)	0.525	0.4796	0.4947
NKV	50.242 (±0.075)	0.247	0.4441	0.4384
Tick	53.963 (±0.016)	0.590	0.516	0.5407
All *flaviviruses*	52.840 (±0.025)	0.529	0.4824	0.4981

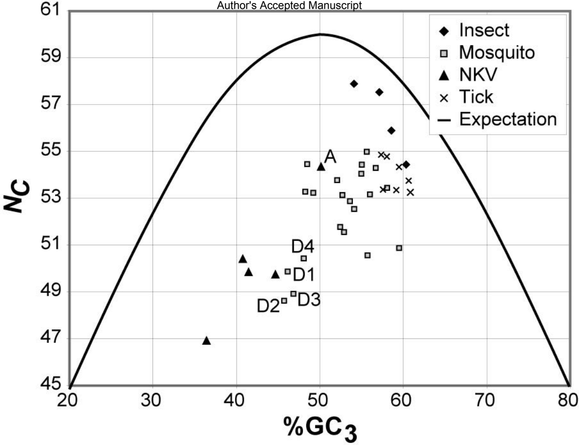
LEGENDS TO FIGURES

Figure 1. Distribution of GC_3 and N_C. The data points labeled “A”, “D1”, “D2”, “D3”, and “D4” identify APOIV, DENV1, DENV2, DENV3 and DENV4, respectively. The solid line shows the expectation of N_C given GC_3 according to the hypothesis that there is no selection and G+C biases at silent sites are due to mutation (Wright, 1990). Tables 1 and 3 can be referenced to identify the individual species in the graph. Grouping based upon host species or vector species is not evident.

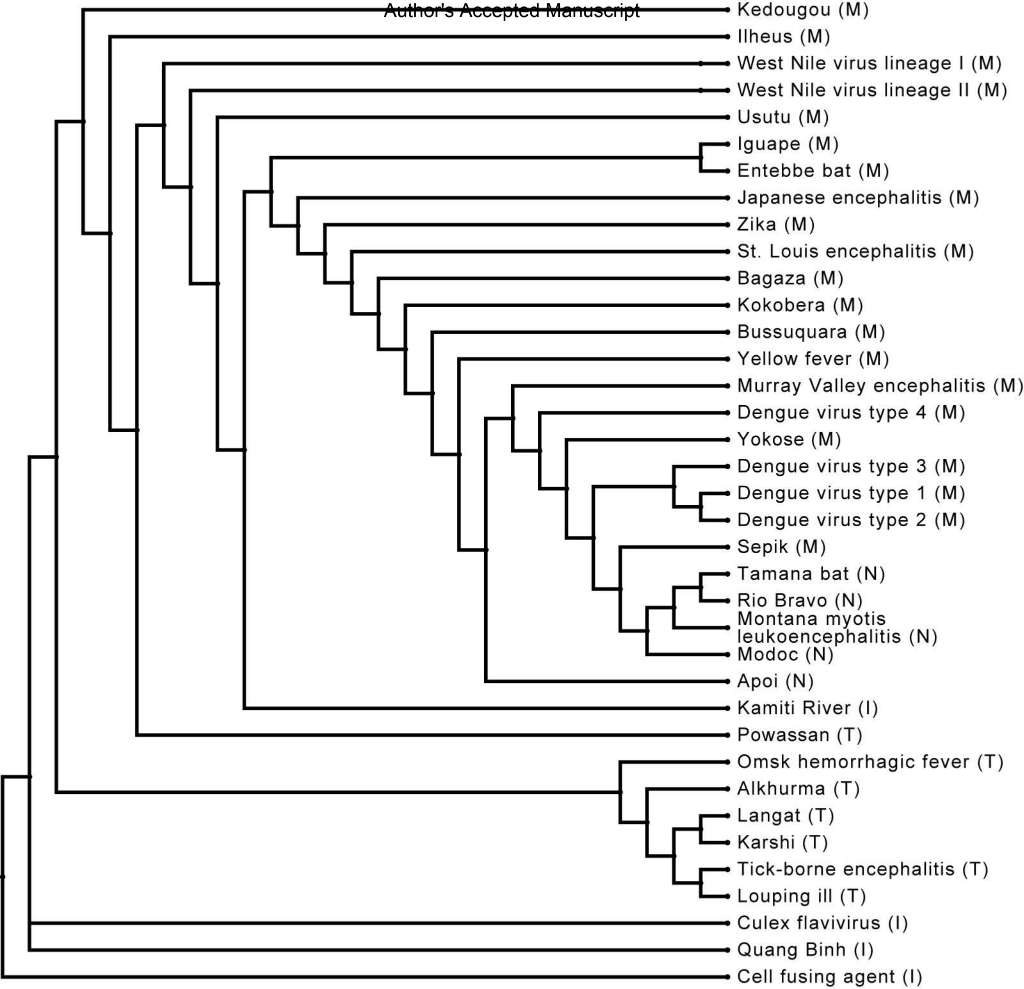
Figure 2. Phylogeny based on codon bias of the complete polyprotein for 37 *flaviviruses*. For each of the viral species considered, the principal vector group (I for insect-specific group, M for mosquito, T for tick, NKV for species with no known vector) is listed in parenthesis.

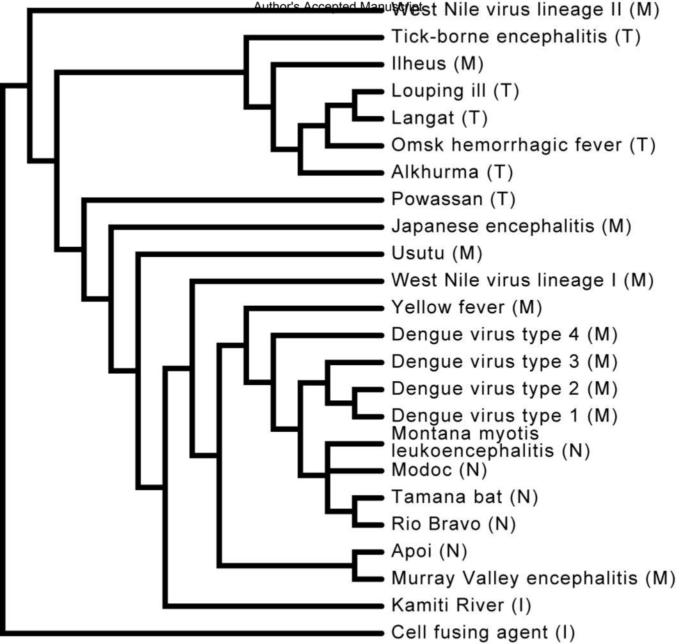
Figure 3. Phylogeny based on NS5 gene codon bias for the 24 species in which gene annotations were included in the RefSeq. For each of the viral species considered, the principal vector group (I for insect-specific group, M for mosquito, T for tick, NKV for species with no known vector) is listed in parenthesis.

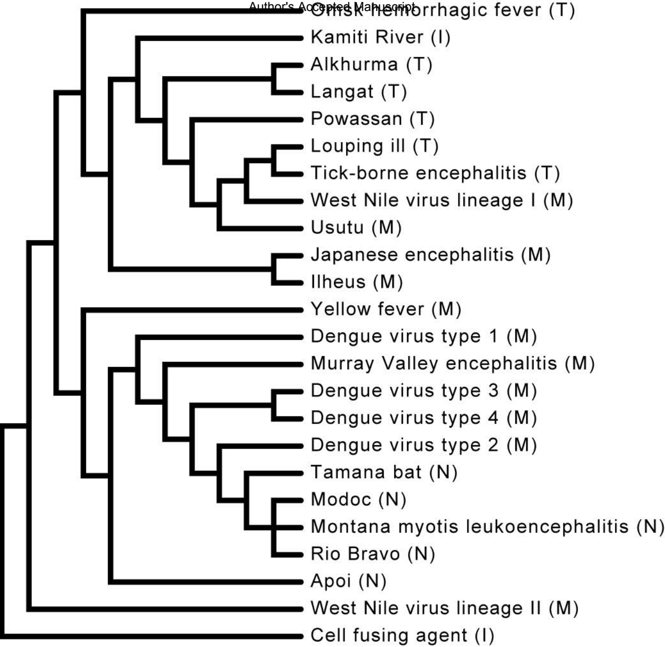
Figure 4. Phylogeny based on NS3 gene codon bias for the 24 species in which gene annotations were included in the RefSeq. For each of the viral species considered, the principal vector group (I for insect-specific group, M for mosquito, T for tick, NKV for species with no known vector) is listed in parenthesis.



61
59
57
55
53
51
49
47
45
20
30
40
50
60
70
80
N_C
%GC$_3$
◆ Insect
▨ Mosquito
▲ NKV
× Tick
— Expectation
A
D4
D1
D2
D3


Kedougou (M)
Ilheus (M)
West Nile virus lineage I (M)
West Nile virus lineage II (M)
Usutu (M)
Iguape (M)
Entebbe bat (M)
Japanese encephalitis (M)
Zika (M)
St. Louis encephalitis (M)
Bagaza (M)
Kokobera (M)
Bussuquara (M)
Yellow fever (M)
Murray Valley encephalitis (M)
Dengue virus type 4 (M)
Yokose (M)
Dengue virus type 3 (M)
Dengue virus type 1 (M)
Dengue virus type 2 (M)
Sepik (M)
Tamana bat (N)
Rio Bravo (N)
Montana myotis leukoencephalitis (N)
Modoc (N)
Apoi (N)
Kamiti River (I)
Powassan (T)
Omsk hemorrhagic fever (T)
Alkhurma (T)
Langat (T)
Karshi (T)
Tick-borne encephalitis (T)
Louping ill (T)
Culex flavivirus (I)
Quang Binh (I)
Cell fusing agent (I)


West Nile virus lineage II (M)
Tick-borne encephalitis (T)
Ilheus (M)
Louping ill (T)
Langat (T)
Omsk hemorrhagic fever (T)
Alkhurma (T)
Powassan (T)
Japanese encephalitis (M)
Usutu (M)
West Nile virus lineage I (M)
Yellow fever (M)
Dengue virus type 4 (M)
Dengue virus type 3 (M)
Dengue virus type 2 (M)
Dengue virus type 1 (M)
Montana myotis leukoencephalitis (N)
Modoc (N)
Tamana bat (N)
Rio Bravo (N)
Apoi (N)
Murray Valley encephalitis (M)
Kamiti River (I)
Cell fusing agent (I)


Omsk hemorrhagic fever (T)
Kamiti River (I)
Alkhurma (T)
Langat (T)
Powassan (T)
Louping ill (T)
Tick-borne encephalitis (T)
West Nile virus lineage I (M)
Usutu (M)
Japanese encephalitis (M)
Ilheus (M)
Yellow fever (M)
Dengue virus type 1 (M)
Murray Valley encephalitis (M)
Dengue virus type 3 (M)
Dengue virus type 4 (M)
Dengue virus type 2 (M)
Tamana bat (N)
Modoc (N)
Montana myotis leukoencephalitis (N)
Rio Bravo (N)
Apoi (N)
West Nile virus lineage II (M)
Cell fusing agent (I)