Phylogenomic characterization of flaviviruses

Phuoc Truong Nguyen

Master's thesis

University of Turku
Department of Biology

Field: Physiology and Genetics
Specialization: Genetics

Credits: 40 ECTS

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

UNIVERSITY OF TURKU
Department of Biology
Phuoc Truong Nguyen
Phylogenomic characterization of flaviviruses
Thesis, 62 pages (4 appendices).
Biology
February, 2019

Background: The occurrences of global viral pandemics have been rising as increased travel between distant countries has introduced previously endemic viruses to new environments. Major contributors to global human hemorrhagic and neurological diseases with high mortality rates include half of the ca. 70 species of the genus *Flavivirus*. The most widespread and well-known flaviviruses are Dengue virus, Japanese encephalitis virus, West Nile virus and Zika virus. Although the transmission routes of major viruses are well-documented and thoroughly researched, the knowledge has been gained from past outbreaks, which has been a limitation in surveillance of novel flaviviruses. Thus, having early information about potential hosts is essential in controlling and preventing viral outbreaks.

Aims: The goal of the master's thesis is to characterize the codon and nucleotide compositions of flaviviruses and to assess a potential use to the identification of putative hosts. This methodology will be utilized to develop a new algorithm capable of identifying optimal hosts through a simple comparative codon usage analysis. This information will be highly valuable to estimate the risk of spread of a virus.

Methods: The genomic characterization of flaviviruses was done with computational biology methods. Computed codon usages were analyzed with clustering methods to identify subgroups of viruses and their optimal hosts. The rationale behind this methodology was that codon usages vary among species and this variability is driven by the virus adaptation to the hosts.

Results: (1) Genotypes of Zika viruses showed distinct codon usage patterns, which linked the origin of American and European virus cases to the Asian genotype. (2) Distinct usage patterns were similarly observed when the methodology was applied to other major flaviviruses. (3) Optimal hosts for mosquito-borne flaviviruses included vertebrates and *Aedes* mosquitos, whereas tick-borne viruses were optimized to ticks. *Aedes* mosquitoes were also optimal for insect-only flaviviruses. *Culex* and *Anopheles* mosquitoes were suboptimal to all groups. Moreover, flaviviruses clustered based on established vector-based classification, host types preferences and phylogeny. The identified hosts were in accordance to previous studies done in field and laboratory.

Conclusions: The proposed methodology based on codon usages is able to estimate hosts for flaviviruses within a close range. The algorithm can be implemented in computationally weak equipment, thus it may be deployed fast and on-site during viral pandemics. In further studies this methodology, with minor modifications, could be utilized to predict putative hosts of other viruses. A scientific article describing the host identification algorithm is under preparation (appendix 4).

Keywords: virus, flavivirus, relative synonymous codon usage (RSCU), codon adaptation index (CAI), algorithm, computational biology, host identification

TURUN YLIOPISTO
Biologian laitos
Phuoc Truong Nguyen
Flavivirusten fylogenominen määritys
Tutkielma, 62 sivua (4 liitettä).
Biologia
Helmikuu, 2019
Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check -järjestelmällä.

Taustatiedot: Maailmanlaajuisten viruspandemioiden määrä ihmisten keskuudessa on kasvanut, koska ihmisten lisääntynyt liikkuminen on mahdollistanut endeemisten virusten leviämisen uusiin ympäristöihin. Noin 70 *Flavivirus*-suvun virusta kuuluu kuolemaan johtavien verenvuoto- ja hermostotautien merkittävimpiin aiheuttajiin. Laajimmalle levinneet ja tunnetuimmat flavivirukset ovat denguevirus, Japanin aivotulehdusvirus, Länsi-Niilin virus ja zikavirus. Virusten leviämismekanismit ovat yleensä selvitetty vasta virusepidemioiden jälkeen. Tiedon puute virusten siirtovektoreista ja isäntäeliöistä rajoittaa uusien flavivirusten jäljittämistä, minkä vuoksi ennakko tieto mahdollisista isäntäeliöistä on välttämätön virusepidemioiden hallinnassa ja torjunnassa.

Päämäärät: Päämääränä on määritellä flavivirusten koodaavan sekvenssin kodoni- ja nukleotidikoostumus sekä hyödyntää näitä mahdollisten isäntäeliöiden tunnistuksessa. Menetelmää hyödynnetään kehitettäessä uutta algoritmia, jolla voitaisiin tunnistaa optimaaliset isäntäeliöt yksinkertaisella vertailevalla kodonienkäyttöanalyysillä.

Menetelmät: Laskennallisen biologian menetelmillä määritettiin flavivirusten kodoni- ja nukleotidikoostumus. Virusten alaryhmät ja optimaaliset isäntäeliöt tunnistettiin lasketuilla kodonien käyttöarvoilla, jotka olivat analysoitu klusterointimenetelmillä. Tunnistusmenetelmä perustui viruslajien erilaisiin kodonienkäyttöprofiileihin, joiden muunteluun vaikuttaa virusten adaptaatio isäntäeliöihin. Tämä tieto on hyvin tärkeää virusten leviämisriskin arvioinnissa.

Tulokset: (1) Zikavirusten eri genotyypeillä havaittiin omanlainen kodonienkäyttöprofiili. Havainto osoitti amerikkalaisten ja eurooppalaisten zikuvirusten kuuluvan Aasian genotyyppiin. (2) Erilaisia käyttöprofiileja havaittiin myös denguevirukselle, Japanin aivotulehdusviruksella ja Länsi-Niilin viruksella. (3) Flavivirusten klusterit ryhmittyivät siirtovektorien, isäntäeliöiden ja fylogenian mukaisesti. Hyttysten levittämien flavivirusten optimaaliset isäntäeliöt olivat selkärankaiset ja *Aedes*-suvun hyttyset, kun taas punkkien levittämät virukset olivat sopeutuneet punkkeihin. *Aedes*-suvun hyttyset olivat myös optimaalisia isäntiä hyönteisiä infektoiville viruksille. *Culex*- ja *Anopheles*-suvun hyttyset eivät olleet ihanteellisia isäntäeliöitä millekään flavivirusryhmälle. Algoritmin tulokset olivat yhteneviä aiempien kenttä- ja laboratoriotutkimusten tulosten kanssa.

Johtopäätelmät: Tutkielmassa kehitetty menetelmä, joka perustuu kodonien käyttöön, pystyy arvioimaan suhteellisen tarkasti flavivirusten isäntäeliöt. Koska algoritmi ei tarvitse paljon laskentatehoa, menetelmää voidaan hyödyntää kenttäolosuhteissa. Jatkokehitettyä menetelmää voitaisiin soveltaa jatkossa muidenkin virusten isäntäeliöiden määrityksessä. Tutkielman tulokset tullaan julkaisemaan tieteellisessä artikkelissa (liite 4).

Asiasanat: flavivirus, synonyymien (samaa aminohappoa koodaavien) kodonien suhteellinen käyttö, kodonien adaptaatioindeksi, algoritmi, laskennallinen biologia, isäntäeliön tunnistus

CONTENTS

1. INTRODUCTION

1.1. Virus epidemics in the recent past

The number of novel human infecting viruses and pandemics have been steadily increasing since the beginning of the 20th century (Woolhouse et al. 2012). The main factors that contribute to the rise of viral epidemics are international traveling, which contributes to the introduction of endemic viruses and their vectors to new environments (Tatem et al. 2006); climate change, which widens the habitable zone of many disease transmitting vector organisms (Githeko et al. 2000); and the rise of human population densities that can increase the likelihood of exposure and transmission of pathogens (Dobson and Carper 1996; Wolfe et al. 2007).

It is estimated that 61% of human infecting pathogens are zoonotic (Taylor et al. 2001), i.e. mainly spread from animals to human. Out of these, vector-borne viruses make up 11% (Taylor et al. 2001). However, novel human infecting pathogens are estimated to be 75% zoonotic and of those, 29% are vector-borne viruses (Taylor et al. 2001). Novel human infecting viruses primarily emerge through the process of initial exposure and the following primary infection (Wolfe et al. 2007). Many of the current viral infectious diseases in humans have originated from zoonotic viruses, such as AIDS (Keele et al. 2006), swine influenza (Smith et al. 2009; Das et al. 2010), Ebola virus disease (Gire et al. 2014), rabies (Badrane and Tordo 2001) and measles (Furuse et al. 2010).

1.2. Genus *Flavivirus*

The name of these arthropod-borne, i.e. arboviruses, originates from the Latin word *flavus* meaning yellow, which stems from the yellow skin pigmentation (jaundice) caused by yellow fever. Flaviviruses can be found on every continent, but species have diverse geographic distributions. Infections from Flaviviruses may cause a wide-range of symptoms including several types of fever (e.g. meningitis, encephalitis, fever and hemorrhagic fever) and other ailments (e.g. arthralgia and rashes). Over half of the ca. 70 species of the genus *Flavivirus* cause illnesses with high mortality rates and are major contributors to global human hemorrhagic and neurological diseases.

Flaviviruses originated right after the last ice age, around 10,000 years ago, from an ancestral non-vector mammal specific virus (Gould et al. 2001, 2003). The serological classification of flaviviruses divides them into two major groups: viruses with and without a vector. The monophyletic groups of vector dependent flaviviruses are based on the type of vector (Kuno et al. 1998). These are 1) mosquito-borne flaviviruses (viruses mainly spread by mosquitoes); 2) tick-borne flaviviruses (viruses mainly spread by ticks); 3) insect-specific or insect-only flaviviruses (viruses that infect only insects); and 4) flaviviruses with unknown vector.

Flaviviruses infect a wide range of vertebrates and hematophagous arthropods (Chambers et al. 1990), mainly mosquitoes (50% of flaviviruses) and ticks (28% of flaviviruses) (Simmonds et al. 2017). As arboviruses, vector-borne flaviviruses have enzootic transmission cycles, which are also known as jungle (sylvatic) cycles. During these cycles, a virus replicates itself within the cell of a reservoir host without causing any apparent harm to the host. Therefore, reservoir hosts tend to be infected multiple times by flaviviruses and exploited for viral reproduction.

1.3. Arthropod vectors of flaviviruses

The most widely distributed flaviviruses are Dengue virus (DENV), Japanese encephalitis virus (JEV), West Nile virus (WNV) and Zika virus (ZKV). They are spread by mosquitoes of the genus *Aedes*, mainly *Aedes aegypti* (yellow fever mosquito) and *A. albopictus* (Asian tiger mosquito). Both species are similar morphologically (dark brown or black with white stripes on their bodies and legs) and are wide-spread around the globe, but mostly in the tropics. Usually, *A. aegypti* is a more suitable vector to spread the virus in human populations because of its ecological habits. Whereas *A. albopictus* prefers to breed, feed and live in rural outdoor areas, *A. aegypti* usually feeds in urban areas within human housings and lays its eggs in small water containers near houses. *A. aegypti* has two distinguishable subspecies, the darker rural *A. aegypti formosus* (Mattingly 1957), and the globally distributed domestic *A. aegypti aegypti*. The domesticated subspecies is more susceptible to Dengue viruses infections enabling it to be a more potent carrier vector (Failloux et al. 2002).

The majority of tick-borne flaviviruses (TBFVs) are transmitted to vertebrates by hard ticks (*Ixodidae*) of the genera *Ixodes*, *Dermacentor* and *Haemaphysalis*. Soft ticks (*Argasidae*) are the primary cause of transmission of TBFVs in seabirds, which rarely infect humans (Estrada-Peña and Jongejan 1999), except when humans go into the natural habitats of soft ticks, e.g. bird nests.

1.4. Flavivirus genome organization and replication

The average size of a spherical flavivirus virion is 50 nm and typically contains three structural proteins: the capsid protein C, the envelope glycoprotein E and the membrane protein M (or its precursor prM). C is a vital capsid protein in the formation of the nucleocapsid, i.e. the unenveloped capsid housing the viral genome. E protein is the most abundant envelope protein. It is dimeric and rod-shaped, and as a hemagglutinin (a glycoprotein that causes the agglutination of red blood cells), participates in both the binding of the surface receptors of a virus and the host cell, and the entry of the virion via endocytosis. The membrane-associated M and prM proteins are used to differentiate virions

that are either mature or immature. If the proteolytic cleavage of it from prM is compromised, it will inhibit the formation of functional viral particles (Amberg and Rice 1999) and therefore decrease the virulence of a virus. The weight of a flavivirus virion consists of approximately 17% of lipids and 9% of glycolipids and glycoproteins, which are derived from host cells during exocytosis.

A flavivirus genome (figure 1) is typically an enveloped positive single-stranded RNA ranging between 9.2–11.0 kilobases in length. The sequence contains a type I cap (m7GpppAmp) and a highly conserved guanine at the 5' end (Wengler et al. 1978), which is unique to the genus (Cleaves and Dubin 1979). The RNA lacks a poly-A tail at the 3' end, although this element has been found in tick-borne encephalitis viruses (Asghar et al. 2014). The RNA strand ends with a conserved cytosine-uracil dinucleotide instead. The coding sequences of flavivirus RNA translates into a polyprotein, from which structural and non-structural proteins are then later cleaved by viral NS2B-NS3 serine proteases. The functions of non-structural viral proteins include controlling the translation and replication of the viral RNA and might be involved in its final packaging into a virion. The differences among flavivirus genomes stem from the varying lengths, nucleotide compositions and the arrangement of the RNA elements. Different RNA element organizations can be particularly observed between mosquito-borne and tick-borne flaviviruses. For example, each has a nonhomologous sequence for RNA cyclization (essential for viral replication) at different genomic positions (Kofler et al. 2006).

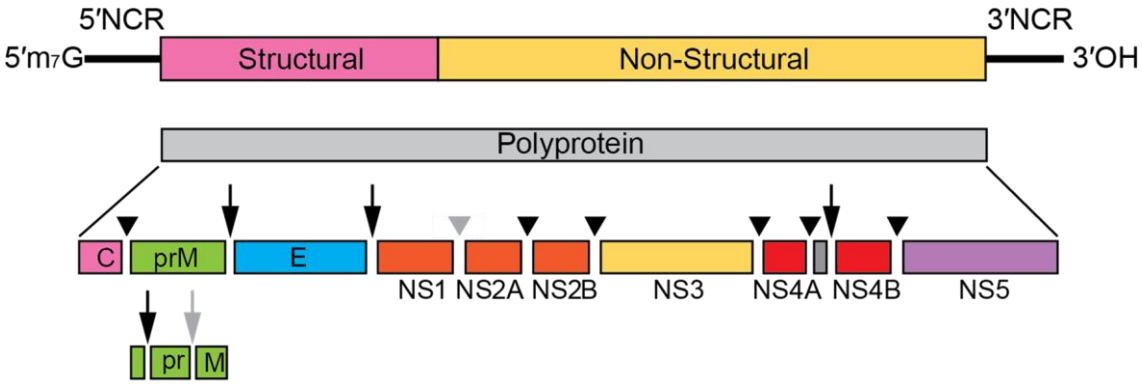

Figure 1. An illustration of a typical flavivirus genome and gene arrangement. The coding sequence for the polyprotein is located between terminal non-coding regions (NCRs). The five-prime cap is marked at the 5' terminal end. The total length of a typical RNA is 9.2–11.0 kilobases. The names of the encoded structural proteins are marked inside boxes and the names of proteases are marked below their respective boxes. Black arrows indicate parts which are processed by signal peptidases, grey triangles by unknown proteases, black triangles by NS2B-3 proteases and grey arrows by Golgi proteases. The image was edited from an original image from ICTV (ICTV 2017; Simmonds et al. 2017) and is free to use under the Creative Commons Attribution 4.0 International License.

A flavivirus genome along with structural proteins encodes seven nonstructural proteins in infected host cells. These are NS1 (46 kDa), NS2A (22 kDa), NS2B (14 kDa), NS3 (70 kDa), NS4A (16 kDa), NS4B (27 kDa) and NS5 (103 kDa). NS1 has an important

role in RNA replication and participates in regulating the activation of the host complement system. It also forms serine protease complexes with NS2B to process the viral polyprotein. NS3 is involved in RNA replication in three aspects: 1) as an RNA helicase, 2) has RNA triphosphatase activity and 3) forms the 5' cap at the end of the strand. NS5 functions as an RNA-dependent RNA polymerase and has methyltransferase activity which modifies the 5' cap. The protein is also the largest and most conserved. Some of the nonstructural proteins induce programmed -1 ribosomal frameshifts, which cause some ribosomes to shift their reading frame during translation by one nucleotide. This introduces transframe fusion proteins (Firth and Atkins 2009), e.g. the protein NS1' that may be involved in the neuroinvasiveness of the virus (Melian et al. 2010).

During an infection, flaviviruses rearrange the perinuclear endoplasmic reticulum (ER) membrane structures to form vesicles where viruses are replicated. The replication process is semi-conservative and includes intermediate RNA molecules in different conformations. The viral RNA first acts as a template for the synthesis of negative-sense complementary RNAs, i.e. replicative intermediates, which in turn are templates for amplifying the number of positive-sense RNAs. These intermediates may be in either single- or double-stranded forms. The translation of open reading frames (ORFs) begins from the start codon AUG.

The first viral particles can be found in ribosome-covered rough ER, in which the virus components are assembled into immature virions. The topology and movement of viral proteins in ER and cytoplasm is controlled by signal and stop-transfer sequences. The virions are then transported by the host cell's secretory system to the cell membrane, from which they are released via exocytosis. The virions mature when the prM protein is removed by furins or a furin-like cellular proteases during the release (Stadler et al. 1997). The host cell releases simultaneously a small noninfectious subviral particle. The host cell might, however, occasionally release viruses with premature virions.

1.5. The importance and challenges of virus surveillance

The most widespread and known flaviviruses (DENV, JEV, WNV and ZKV) are responsible of millions of new infections every year: DENVs cause between 50 to 390 million cases of dengue fever annually (Rigau-Pérez et al. 1998; Bhatt et al. 2013); JEVs are responsible of approximately 65,000 annual cases of encephalitis (Campbell et al. 2011); in 2017 there were around 2,000 and 200 cases of WNVs in the United States (CDC 2017) Europe and Israel (ECDC 2017) respectively; and ZKVs have over 220,000 confirmed cases in the Americas since 2015 (PAHO). Each major flavivirus was capable of spreading nearly worldwide within a couple of decades (WHO 2018; Sejvar 2003; Wang and Liang 2015; Kindhauser et al. 2016). For example, ZKV, which was originally sequenced from Rhesus macaques (*Macaca mulatta*) in Uganda in 1947 (Dick et al. 1952),

became well-known due to its relatively fast pandemic outbreak in South America. The World Health Organization (WHO) received the first reports of South American ZKV cases and of their connection to microcephaly in newborn babies from the Brazilian Ministry of Health in 2015 (Kindhauser et al. 2016). Moreover, the spread of flaviviruses is also responsible of significant economic damage, because they are pathogenic to several animals, both domestic (turkeys, pigs, horses, sheep and dogs) and wild (grouse and muskrats).

In general, flaviviruses cause silent infections, i.e. cause no diseases in their primary hosts or vectors (not to be confused with latent infections), which is why determining the transmission pathways of viruses can be difficult. Thus, it is a challenge to assess in advance the risk of a flavivirus outbreak. This causes problems especially when the viruses enter into urban transmission cycles and begin to spread within human populations after being transmitted by bridge vectors, also known as secondary vectors. An example of a bridge vector is *Culex pipiens* (Hamer et al. 2008), which although mainly ornithophilic, also opportunistically feeds on humans when available and simultaneously spreads WNV (Hamer et al. 2009). Usually flavivirus infections are insignificant for the health of reservoir hosts. However, the infections can manifest severe and even lethal diseases on secondary dead-end hosts, which is why many flaviviruses can remain silent until the habitats of the viruses and new potential hosts overlap. In addition to external factors, viral prediction methods tend to be uncertain due to the many factors relating to the high variability within viruses. This stems from their high mutation rates (Sanjuán et al. 2010), common to all single-stranded RNA viruses, such as flaviviruses, because of low-fidelity polymerases, lack of proofreading mechanisms and rapid amplification of viral genomes within host cells (Sanjuán and Domingo-Calap 2016). Another problem for host prediction is the uncertainty when determining the host organisms. For example, flaviviruses have been repeatedly isolated from animals that are not their primary hosts, which is most likely due to transmission through blood meals between hematophagous vectors and hosts (Kuno 2007). As a result of this, there are organisms incorrectly labeled as hosts leading to erroneous classification of new pathogenic viruses, reproduction cycles and transmission routes. Moreover, these inaccurate findings may lead to imprecise surveillance, and thus, affect the means of efficiently containing viral outbreaks.

1.6. Aims

The aim of this master's thesis is to characterize the genome of flaviviruses and to assess the potential use of computational biology methods, based on the relative use of synonymous codons, to identify their optimal and putative hosts. Due to the prevalence of several flaviviruses and the recent rise of awareness for ZKV, it is relevant to develop

bioinformatics tools to be used as an early-alarm surveillance system. By understanding the evolution and the genomic properties of flaviviruses, genomic markers could be identified to manage and prevent future viral outbreaks. This master's thesis is divided in three phases:

1) A pilot study to test, whether a quantitative characterization based on the relative synonymous codon usage (RSCU) is a valid and specific genomic marker. This methodology is tested with genomes from all currently available ZKV (appendix 2 table 2). The results are then compared with those obtained from the literature.

2) The second phase consists in applying the methodology developed in the first phase to characterize the genomic variability of all fully sequenced flaviviruses. The genomes of flaviviruses are analyzed based on several well-stablished metrics, such as the RSCU, the percentage of guanine and cytosine at the third position of the codon (GC3) and the Codon Adaptation Index (CAI).

3) The third phase implements a novel algorithm, based on a normalized version of the classical CAI (nCAI), to estimate optimal and potential hosts. The results of putative virus-hosts associations are then interpreted based on a multifactorial analysis.

2. MATERIALS AND METHODS

2.1. Data collection

Viral sequences were obtained from public databases via FTP (see below) and stored in servers from the IT Center for Science (CSC 2017).

ftp.ncbi.nlm.nih.gov/genomes/Viruses
ftp.ncbi.nlm.nih.gov/genomes/refseq/viral
ftp.ncbi.nlm.nih.gov/refseq/release/viral

Genomic and protein sequences of flaviviruses were updated regularly when new data was available.

2.1.1. Genome sequences

The classification of flaviviruses and sequences, including amino acids (AAs) and coding sequences (CDSs), were obtained from the databases of the National Center for Biotechnology Information (NCBI 2017; NCBI Resource Coordinators 2018), which is an internationally recognized resource for gathering, storing and analyzing information relating to molecular biology, biochemistry and genetics. Protein sequences and CDSs for DENVs, WNVs and ZKVs were obtained from the NCBI's Virus Variation Resource (Virus Variation Resource 2017; Hatcher et al. 2017) and sequences for JEVs were obtained from the NCBI's Nucleotide database (Nucleotide 2017) via E-utilities (Sayers 2008) (table 1).

Table 1. The number of complete coding sequences (CDSs) and amino acid sequences (AA) for each major mosquito-borne flavivirus used for phylogenomic analyses and the obtention. Sequence data of Dengue viruses, West Nile viruses and Zika viruses were obtained from the Virus Variation Resource (Virus Variation Resource 2017; Hatcher et al. 2017) and data of Japanese encephalitis viruses were obtained from the NCBI's Nucleotide database (Nucleotide 2017). The complete list of sequences used for the ZKV pilot study can be found in appendix 2 table 2.

Virus	Abbreviation	CDS/AA	Date
Dengue	DENV	4,865	25.4.2018
Japanese encephalitis	JEV	297	25.4.2018
West Nile	WNV	1,619	25.4.2018
Zika	ZKV	494	25.4.2018
Zika (pilot study)	ZKV	362	13.6.2016

2.1.2. Classification of flaviviruses

Data about subgroups within DENVs, JEVs, WNVs and ZKVs (table 2) was acquired from Virus Variation Resource (Virus Variation Resource 2017; Hatcher et al. 2017), Virus Pathogen Database and Analysis Resource (ViPR 2017; Pickett et al. 2012), and in few cases directly from the scientific literature. DENVs were divided into four distinct

groups according to their serotypes (WHO 2018), i.e. based on their distinct compositions or patterns of surface antigens. A fifth genotype has been reported by (Normile 2013), but it was not included due to the lack of sequences at NCBI and ViPR databases. WNVs were divided into ten lineages (Pauli et al. 2013), of which the lineages 1A, 1B (also known as Kunjin virus) and 2 are well characterized. The other lineages included were lineage 4 (Lvov et al. 2004), lineage 5 (Bondre et al. 2007), lineage 7 (Vazquez et al. 2010; Pauli et al. 2013), lineage 8 (Fall et al. 2014) and lineage 9/4c (Pachler et al. 2014). Lineage 6 (Bowen et al. 1970; Poidinger et al. 1996) was excluded from this thesis due to no available CDSs. JEVs were categorized into five genotypes (Chen et al. 1990, 1992; Li et al. 2011; Mohammed et al. 2011) and ZKVs were classified into three distinct genotypes based on phylogenetic studies, which were East African, West African and Asian genotype (Ramaiah et al. 2017).

Table 2. Classification of Dengue viruse (DENV), Japanese encephalitis virus (JEV), West Nile virus (WNV) and Zika virus (ZKV). The table also includes the number and the distribution of complete coding sequences (CDSs) among groups.

DENV		JEV		WNV		ZKV	
Serotype	**CDS**	**Genotype**	**CDS**	**Lineage**	**CDS**	**Genotype**	**CDS**
1	2,043	I	109	1A	1,467	East African	79
2	1,445	II	1	1B	48	West African	7
3	1,031	III	183	2	91	Asian	408
4	346	IV	1	3	1		
		V	3	4	4		
				5	5		
				7	1		
				8	1		
				9/4c	1		

Flaviviruses are generally categorized into four monophyletic subgroups depending on arthropod vectors. These are mosquito-borne flaviviruses (MBFVs), tick-borne flaviviruses (TBFVs), insect-only flaviviruses (IOFVs) and flaviviruses with an unknown vector (UVFVs) (Simmonds et al. 2017). These groups were used as a guidance to identify putative hosts. Moreover, flaviviruses were also analyzed based in a much broader (partially overlapping) classification: vertebrates, mosquitoes and ticks. MBFVs, TBFVs and UVFVs have replicative cycles, in which they infect their primary vertebrate host through an arthropod vector. A paraphyletic subgroup of MBFVs (includes Chaoyang virus, Ilomantsi virus, Lammi virus, Nounané virus and Donggang virus) spreads exclusively within mosquitoes (Blitvich and Firth 2015). In this master's thesis, this last group was called dual-host insect-only flaviviruses (dhIOFVs), however, alternative names have been suggested in the literature, e.g. dual-host affiliated insect-specific flaviviruses (Blitvich and Firth 2015) and MBFV-related viruses (Huhtamo et al. 2014).

2.1.3. Codon usage reference tables

Codon usage tables (CUTs) from the potential hosts were obtained from the Codon Usage Database (last release is from June 15th 2007) (Codon Usage Database 2017; Nakamura et al. 2000) when available or calculated using a computer script, which utilizes the same procedure. The Codon Usage Database constructed CUTs based on genetic sequences from GenBank (GenBank 2017), a comprehensive database that contains nucleotide sequences from approximately 260,000 organisms (Benson et al. 2013).

A list of 16 potential hosts were analyzed in this master's thesis, including vertebrates (mammals, birds, reptiles and amphibians) and arthropods (mosquitoes and a tick) (appendix 2 table 1). To ensure that the CUTs were as reliable as possible, only animals with at least 10,000 CDSs were included in the analysis, with the exceptions of the wild boar (*Sus scrofa*) with 2,953 CDSs and the red junglefowl (*Gallus gallus*) with 6,017 CDSs, due to major interest.

2.2. Characterization of genomic composition

An analysis of the codon usage and the GC3 was used to characterize the genomic composition of the four most common flaviviruses (DENV, JEV, WNV and ZKV). Composition of nucleotides in the third codon position (%N3) vary because of a bias towards certain codons during translation. The Relative Synonymous Codon Usage (RSCU) and the Codon Adaptation Index (CAI) were calculated using the local version of the software program CAIcal (version 1.4) (Puigbò et al. 2008a, 2008b). Because viruses use the translational machinery of the host cell, codon usages of viruses tend to have codon frequencies that mirror those from their hosts (Bahir et al. 2009). However, in certain cases, viruses may use the opposite strategy to hide from the host's defense mechanisms (Mossadegh et al. 2004; Zhou et al. 1999; Cid-Arregui et al. 2003; Karlin et al. 1990; Zhou et al. 2012; Puigbò et al. 2010).

RSCU describes the preferential use for a synonymous codon over another and it is calculated by dividing the observed number of the codon by its expected frequency (Sharp and Li 1986). The expected frequency is based on the assumption that codons for AAs are used equally (Sharp et al. 1986). However, in certain genes, e.g. highly expressed ones, there is a strong bias towards certain codons (Post et al. 1979; Sharp et al. 1986; Puigbò et al. 2008c). These usage patterns in genes are generally the same within an organism, although the amount of bias varies in different genes (Sharp and Li 1986). The RSCU was calculated with equation 1:

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}} \quad (1)$$

In this, x_{ij} is the number of a codon (j) for an observed AA (i) occurring in a gene, and n_i is the number of synonymous codons (1–6) encoding the i. The total number of codons is 64.

The CAI is a numerical value that quantifies the amount of bias towards the use of certain codons in a gene. The CAI can be used to estimate expression levels of a gene, to compare the codon usage among organisms and to assess the success of heterologous expression of a gene in a host organism. Moreover, codon usage biases can be used to study a group of highly expressed genes (Puigbò et al. 2008c), the effect of translational selection (Puigbò et al. 2007), and potential open reading frames of a gene (Sharp and Li 1987, 1986; Puigbò et al. 2008a). In this master's thesis, a variant of the CAI was used to determine the degree of adaptation of viral genomes to putative hosts.

To determine the CAI in a gene, a reference table of RSCU values of the sequence needed to be calculated. The table was constructed by calculating the relative adaptiveness of a codon (w_{ij}):

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{imax}} \quad (2)$$

In equation 2, the value for relative usage for a codon encoding an AA ($RSCU_{ij}$) is divided by the same value for the optimal codon ($RSCU_{imax}$).

The CAI can then be calculated using equation 3 based on the geometric mean of w_{ij} values for each codon in the gene:

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{\frac{1}{L}} \quad (3)$$

L is the number of codons, and w_k is the w value for the codon (k). CAI is calculated from 59 synonymous codons in the standard genetic code, i.e. excluding the stop codons (UAA, UAG and UGA), the start codon (AUG) and the codon that encodes for the AA tryptophan (UGG). CAI can alternatively be calculated more accurately with equation 4:

$$CAI = exp \frac{1}{L} \sum_{k=1}^{L} \ln w_k \quad (4)$$

CAI can also be calculated from a CUT, in which case equation 5 should be used:

$$CAI = exp \frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{n_i} x_{ij} \ln w_{ij} \quad (5)$$

While the effects of the length of a gene (L) are minimal, short sequences may introduce variability and therefore give inaccurate results.

2.3. General statistical methods

Statistical methods used in this thesis included root-mean-square deviation (RMSD) and determination coefficient R squared (R^2).

RMSD is a general method in statistics to measure the distance between multiple datasets to identify the outliers. This is usually done by comparing observed data to expected values by calculating the square root of the mean of the deviations squared (Barnston 1992). In this master's thesis, the RMSD was used to calculate the distance between viral RSCU matrices from different geographic locations (x_i and y_i) with equation 6:

$$RMSD = \sqrt{\frac{\sum (x_i - y_i)^2}{w}} \quad (6)$$

In this equation, w is the sample size, i.e. the 59 synonymous codons. Low RMSD scores indicate that there is less variation between codons in different data sets. The RMSD can be used to identify outliers, however there are limitations if the deviations are too small.

The determination coefficient R squared (R^2) is used to determine the linear relationship between different datasets. It tells how much the x variables explain the variance in y, i.e. how much a model explains the observed variability of the data in relation to its mean values. The coefficient value ranges from 0 (no correlation between datasets) to 1 (perfect correlation between datasets). However, while R^2 is an indication of statistical significance, it does not give information about the quality of the dataset or the model.

2.4. Host identification

A modified version of the CAI was used to identify putative and optimal hosts. This normalized CAI (nCAI) value was the result of dividing the standard CAI values by CAI values obtained from own reference genome as in equation 7:

$$nCAI = \frac{CAI_h}{CAI_s} \quad (7)$$

Host CAI values (CAI_h) were calculated using the program CAIcal (Puigbò et al. 2008a) from virus CDSs and CUTs from known and putative host organisms (appendix 2 table 1). Self CAI value (CAI_s) for each virus was calculated similarly but using the codon usage of the analyzed virus as a reference instead. The CAI value is 1, if the CDS of a virus is compared to its own CUT, because the frequency of codons is identical to the CDS if both are from the same organism. An advantage of the nCAI was that it allowed the comparison among different viruses. The nCAI equaled 1 when the codon usage between a virus and a host organism mirrored each other perfectly. Thus, the closer the nCAI was to 1, the more optimized the virus was to a host and the higher the putative risk of infection. A low level of virus-host codon usage adaptation (underoptimization) was when the nCAI was below 1, and alternatively, a high level of adaptation (overoptimization) was when the value was above 1. Because the range of optimal nCAI values had not been previously established, for this thesis, optimal hosts were considered to be within the conservative range between 0.95 and 1.05.

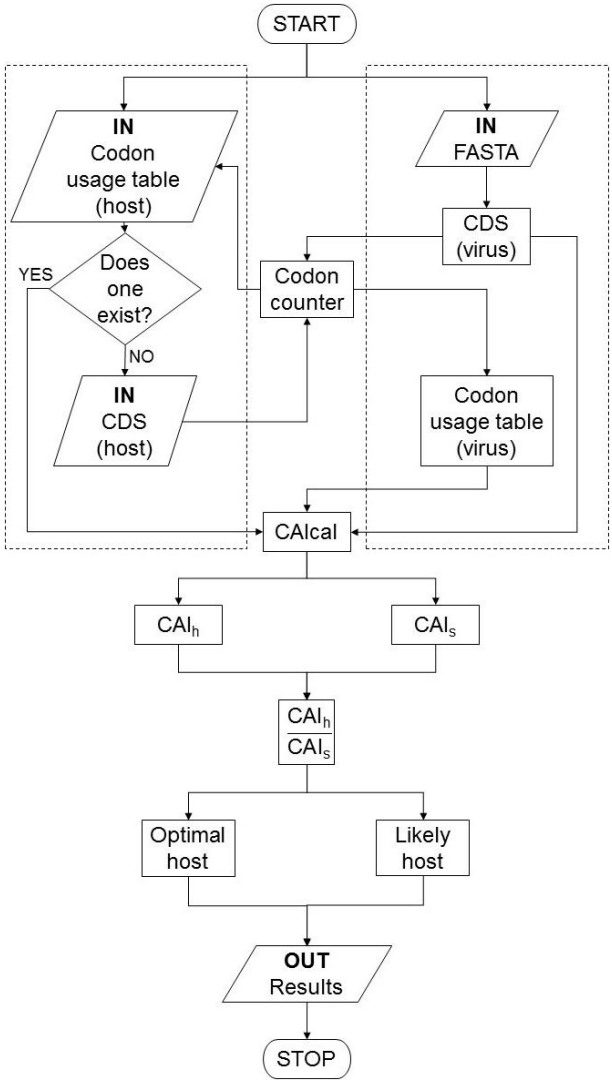

Figure 2. Algorithm to identify optimal hosts based on codon adaptation index (CAI) values. To compute normalized CAI (nCAI) values, it requires CAI values for the host (CAI_h) and the control CAI values for the virus itself (CAI_s). CAI_h is calculated with the coding sequences of viruses and the codon usage tables of hosts (dashed box on the right), while CAI_s is calculated similarly, but with the reference tables of viruses themselves instead (dashed box on the right). The codon usage tables are calculated with computer scripts, and the CAI values are computed with CAIcal (Puigbò et al. 2008a). CAI_h is then divided by CAI_s, and optimal and likely hosts are then determined based on the resulting nCAI values.

The optimal host identification algorithm (figure 2) was run with available CDSs of currently identified flaviviruses (N = 94). The putative hosts (appendix 1 table 1) were chosen based on the current information provided by Virus-Host Database (Virus-Host DB 2017; Mihara et al. 2016), a comprehensive resource for the relationships between viruses and their hosts, and reported cases from the scientific literature. While the database compiles information from several well-known sources, such as Genbank, NCBI and UniProt, the organisms listed may not be always the actual hosts for a virus. This is due to the possibility of viruses being sequenced and subsequently “found” in organisms that may not be hosts at all.

2.5. Cluster analysis

2.5.1. UPGMA

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a method in numerical taxonomy to infer phylogeny between organisms (Sokal and Michener 1958). It is based on agglomerative hierarchical clustering, which builds a dendrogram starting from the bottom, i.e. from smaller clusters, which consist of single observations, and then pairing the most identical clusters to form ever larger clusters. The process starts with a distance matrix, which is a matrix of pair-wise distance values between two elements. By combining elements with similar distances, these form larger composite clusters, which are then treated as one element in the next clustering. The distance values between the new composites are then calculated using the arithmetic averages of the distances between a composite and the data points in another composite cluster. Because all distance values contribute equally when computing new values, UPGMA is called “unweighted”. The more important factor affecting the averages is the number of taxa in a cluster. This differs from Weighted Pair Group Method with Arithmetic Mean (WPGMA), in which the distances are the averages between individual clusters.

In the ZKV pilot study (appendix 2 figure 1), a dendrogram of ZKVs was done with DendroUPGMA (Garcia-Vallvé and Puigbò 2002) using RSCU values of all three genotypes (East African, West African and Asian). The tree construction process begun with obtaining all available CDSs of ZKVs from the Virus Variation Resource (Virus Variation Resource 2017; Hatcher et al. 2017), after which the RSCU was computed with CAIcal (Puigbò et al. 2008a) for each sequence. The results were compiled into a table, which was then analyzed with DendroUPGMA. The workflow is summarized in figure 3. It is important to note, that because the clustering of viruses was determined based on the similarity of RSCU values, the dendrogram computed with UPGMA did not necessarily reflect actual phylogenetic relationships.

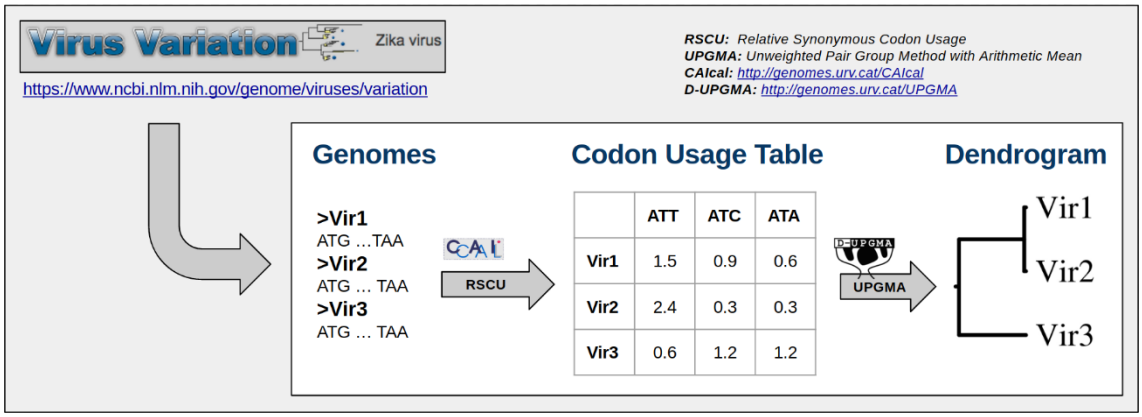

Figure 3. Pipeline showing the steps to reconstruct a dendrogram from relative synonymous codon usage (RSCU) values. Complete coding sequences of Zika viruses were first obtained from Virus Variation Resource database Resource (Virus Variation Resource 2017; Hatcher et al. 2017). The codon usage was then calculated using CAIcal (Puigbò et al. 2008a) and the results put in a table. Finally, a dendrogram was computed from the RSCU values using DendroUPGMA (Garcia-Vallvé and Puigbò 2002). Original figure can be found in appendix 2 figure 1.

2.5.2. K-means

K-means clustering was used in this thesis to evaluate the representativeness of reference sequences of DENVs, JEVs, WNVs and ZKVs. It is an unsupervised method to identify groups within data that does not have defined categories. It was first proposed in the 1950s (Steinhaus 1956), although its present form was later published in the 1980s (Lloyd 1982). The name of the method was popularized in the 1960s (MacQueen 1967).

The k-means algorithm iteratively cycles through two phases: In the first phase, the algorithm first assigns each data point to a predetermined or a random number of groups represented by the variable K. The data points are then clustered based on the similarity of a value, pattern or any other factor. In the second phase, the algorithm calculates a new centroid based on the average of all data points in a formed cluster. The cycle repeats until the location of the centroid and the assigned data points in a cluster do not change.

The clusters are computed based on the shortest Euclidean distances between data points and their respective centroids. These clusters are then labeled based on their computed centroids and the names can be applied to the original data. The centroid weights can finally be examined to qualitatively identify the clustering factor.

There are many benefits to k-means clustering method. It produces results that are easy to interpret, it can be adjusted if there are problems during the analysis, and it can divide data sets efficiently. It also requires less computational resources compared to other clustering methods, e.g. dendrograms. However, the results may vary depending on the initial K. Therefore, to produce accurate results, the inputted K must be appropriate to the size and distribution of the dataset. Additionally, if the analyzed data is too homogeneous (anisotropic), the produced k-means clusters may not reflect actual groups.

2.6. Multifactorial analyses

Multifactorial analysis (MFA) is a method to study data tables with observations or variables that may be either numerical (quantitative) or categorical (qualitative). Based on these, the data can be grouped and the amount of correlation between groups can be determined by analyzing their relationships. Elements on a table can belong into several groups. MFA consists of two steps: the data is first analyzed with a principal component analysis (PCA) (Pearson 1901; Hotelling 1933), which is an extension of MFA. After this, the results are normalized by dividing each with the first value of the PCA. The normalized values are then collected into a new table in the second step and run through another PCA without normalization of values. The final PCA gives factor scores to the observations and loadings for the variables.

Correspondence analysis (CA) (Hirschfeld and Wishart 1935) is a multifactorial method to measure the correlation or relationship between multiple variables in a contingency table, which is commonly a two-way table with two variables. The method plots the rows and columns of the table as dots. Additionally, it can show which of the column variables contribute the most to the similarity or difference between the variables.

In this thesis, to determine the evolutionary relationship between different subgroups of DENVs, JEVs, WNVs and ZKVs (table 2), a simple CA was performed using the R package “ca” (version 0.70). In a simple CA, the previously calculated CAI and RSCU values were analyzed as separate variables. The results were then plotted with the R package “ggplot2” (version 2.2.1).

2.7. Phylogenetic analyses

Phylogenetic tree of flaviviruses (appendix 3 figure 1) was built using available reference sequences of all flaviviruses and if not available, the sequences were substituted with other representative sequences. The sequence data was obtained from NCBI Reference Sequence Database (RefSeq 2017; O’Leary et al. 2016). The aim of RefSeq is to create a comprehensive database of essential DNA, RNA and protein sequences of eukaryotes, prokaryotes and viruses. The database currently has sequences ranging from single genes to complete genomes from over 55,000 organisms (O’Leary et al. 2016). The mosquito-borne flaviviruses were selected based on known classification (Kuno et al. 1998; Simmonds et al. 2017). Tree of representative hosts was constructed based on NCBI Taxonomy database (NCBI Taxonomy browser 2018; Sayers et al. 2009; Federhen 2012; Benson et al. 2013), which aims to incorporate phylogenetic and taxonomic information from published literature, web-based databases, and the advice of sequence submitters and taxonomy experts. However, the database notes that it is not an authority

in phylogeny or taxonomy. The hosts were chosen based on recorded cases in literature (appendix 1 table 1).

Separate phylogenetic trees for DENVs, JEVs, WNVs, ZKVs and for all flaviviruses (appendix 1 table 1) were made by performing a multiple sequence alignment (MSA) with the AA sequences using MUSCLE (Edgar 2004). MUSCLE is one of the most used and cited MSA software, able to create alignments for hundreds of nucleotide or AA sequences in seconds. Usually, the main aim of MSA is to determine the level of homology between analyzed sequences, which can then be used to infer phylogeny. In this thesis, protein sequences were used to build phylogenetic trees (instead of nucleotides) primarily because AAs provided twenty variables to analyze compared to four nucleotides, and thus they represented evolutionary relationships more accurately (Hall 2005).

The alignment data was then analyzed with FastTree (version 2.1.10) (Price et al. 2009) and RAxML (version 8.1.3) (Stamatakis 2014) to obtain maximum-likelihood phylogenetic trees. FastTree is a quick open-source tool to infer phylogenetic relationships from distance-matrix values using the “minimum-evolution” principle, which attempts to build phylogenetic trees based on a topology with the least amount of evolution, or in which the sum of the branch lengths is the lowest. It searches a starting tree using heuristic neighbor joining method. The neighbor joining begins from an initial distance dataset from which the neighbors, i.e. operational taxonomic units (OTUs), are paired and this results in a certain amount of total branch length in a star-shaped tree. At each pairing or clustering phase, the total amount of branch length is minimized, and at the end gives a parsimonious tree (Saitou and Nei 1987; Studier and Keppler 1988). Then there is a minimum-evolution phase, during which the software attempts to reduce the length of the tree by using both nearest-neighbor interchange (NNI) and subtree-prune-regraft (SPR) rearrangement methods. NNI switches around branches (subtrees) of a main tree, while SPR removes each branch individually from the main tree and inserts it to a different place (Felsenstein 2004). Both rearrangement strategies produce multiple hypothetical trees, from which the most optimal is then chosen. The topology and branch lengths of the tree is then further optimized by additional NNIs. RAxML (Randomized Axelerated Maximum Likelihood), an open-source program, utilizes the maximum-likelihood principle, which is more accurate but time-consuming and computationally demanding method than FastTree.

The robustness of the phylogenetic trees was estimated based on a method called Bootstrap (Efron 1979). Bootstrapping is a resampling method to measure the accuracy of statistical analyses by estimating the variance, i.e. uncertainty of used data. In phylogenetics, bootstrap values are used to indicate the probability of a tree being accurate. This

is accomplished by computing how many times a clade appears when reanalyzing samples from the initial dataset (Felsenstein 1985). Both applied tree building software are capable of calculating bootstrap values, but due to the high number of sequences, and therefore the high amount of computational power required, only FastTree was used to calculate bootstraps.

The phylogenetic trees were visualized using iTOL, the Interactive Tree Of Life (iTOL 2017; Letunic and Bork 2016). The iTOL (Version 4.0.3) is a web-based tool to edit and display large phylogenetic trees with up to 100,000 leaves. The tool was used to reroot the trees to an outgroup, annotate all virus genomes and display bootstrap values for every branch.

2.8. Programming skills

The manipulation and analyses of large genomic datasets was primarily done using the scripting language Perl (version 5.26.1) and the R package (version 3.4.4). Both are viable option for bioinformatics, because of their multitude of publicly available modules and packages, which can be utilized in many fields of biology. As an undergraduate student, I did not have prior experience in Perl, R or any other programming language, thus required additional training for the completion of this thesis.

The data manipulation with Perl included data collection from NCBI database servers, parsing of the data, and creation of datasets with relevant genomic information to use in further analyses. These were mainly accomplished with self-coded scripts, which were compiled into bioinformatics pipelines. Most of the bioinformatics analyses, e.g. sequence analyses and phylogenetic tree constructions, were done with publicly available software, while applied statistics, multifactorial analyses and data visualizations were done with R packages.

All programming and bioinformatics work were done in a local Linux operating system (Ubuntu 16.04). Data storage and computationally demanding analyses, e.g. protein alignments and phylogenetic tree constructions, were done in CSC servers (CSC 2017). Working in these environments required also additional training.

3. RESULTS

3.1. Pilot study

The validity of the quantitative characterization method based on the RSCUs of ZKVs was assessed by constructing a dendrogram from RSCU values and comparing it to established literature. The aim was to test, whether each genotype had specific codon usages and whether these could have been used to infer phylogeny. The calculated averages of RSCUs displayed patterns in codon usage that may have been unique to the virus (appendix 2 table 3). In addition, these values showed genotype specific codon frequencies (appendix 2 table 3) and variations of GC at third codon position (table 3). The dendrograms constructed from RSCU values (figure 5) showed similarities between viruses from different geographic locations and from different genotypes. The dendrograms and CA plots distinguished three genotypes (East African, West African and Asian), which matched the phylogenetic classification and revealed the origin of ZKV sequences from the Americas and Europe. Quantitative analyses with R^2 (figure 7) and RMSD (table 4) showed similar results to the phylogenic tree. Furthermore, the dendrograms (figure 5) and RMSD values (table 4) supported the hypothesis that the Asian genotypes differentiated early in its evolutionary history from an ancestral African genotype.

3.1.1. Codon usage patterns

The RSCU averages showed that there were specific regional preferences for codon usages with ZKVs from East Africa (average SD = 0.033), West Africa (average SD = 0.019) and Asia (average SD = 0.032) (appendix 2 table 3). Certain codons were more frequent in African genotypes, while others were more common in the Asian genotype. However, these differences were minor. The codon usage bias was on average more pronounced in AAs that are encoded with more than two codons. Overall, the most abundant codon was on average AGA (encodes arginine), and the least abundant was CGA (encodes arginine).

Multiple synonymous codons (two or more) encoding the same AA showed high variation in usage biases (appendix 2 table 3). The difference between the higher and lower average RSCU value for these codons was 2.351. The most frequent codon among all ZKVs was on average AGA (encodes arginine) and the least was CGA (encodes arginine). While AGA was most common in all genotypes, the most uncommon codons differed between the genotypes. CCG (encodes proline) was the least frequently used in East African genotype, TTA (encodes leucine) was the least frequent in West African genotype, and CGA (encodes arginine) was the most uncommon codon in Asian genotype.

Among the AAs encoded by two alternatives codons, there was on average less variance than in those that are encoded by a higher number of alternative codons (appendix 2 table 3). The range between higher and lower average RSCUs was 0.67 for these codons. The most frequently used codon for those AAs encoded by only two codons in ZKVs was on average AAC, which (encodes asparagine), whereas the least frequently used codon was AAT (encodes asparagine). East African and Asian genotypes had the same most and least used codons as ZKVs overall. However, while the least common codon in West African genotype was the same as with other genotypes, the most common codon was AAG (encodes lysine).

The comparison between recently isolated American ZKV sequences and the Asian genotype revealed that ZKVs in the Americas were most similar to viruses of the Asian genotype (appendix 2 table 3). There was no significant difference in RSCU or %N3 between them, although there were minor variations in codon usage. Certain codons were more frequent in American ZKVs, e.g. CAA (encodes glutamine), TTG (encodes leucine) and TGC (encodes cysteine). Certain codons were also used less frequently in American ZKVs, e.g. CTT (encodes leucine), CAG (encodes glutamine) and TGT (encodes cysteine). These results indicated that there might be certain selection pressure that leads to changes in the codon frequencies in ZKVs from the Americas, although in CA Asian and American ZKVs cluster together and are practically indistinguishable.

The CA based on RSCU values and nucleotides at the third codon position showed the same results (figure 4). The results based on RSCU values clearly split apart the African and Asian genotypes (figure 4A), explained mostly by the CGA and TCG. These codons encode for the AAs arginine and serine, which are both encoded with six codons. These codons had a very low RSCU value, which indicated that the low frequency of CGA and TCG were specific to the African genotypes. The loads of the CA showed that most of the codons, that clustered the Asian genotype apart from the African genotypes, had either C or G in the third codon position (ca. 61% of all codons). Similarly, the African genotypes clustered due to having mostly codons that ended with either A or T (ca. 61% of all codons).

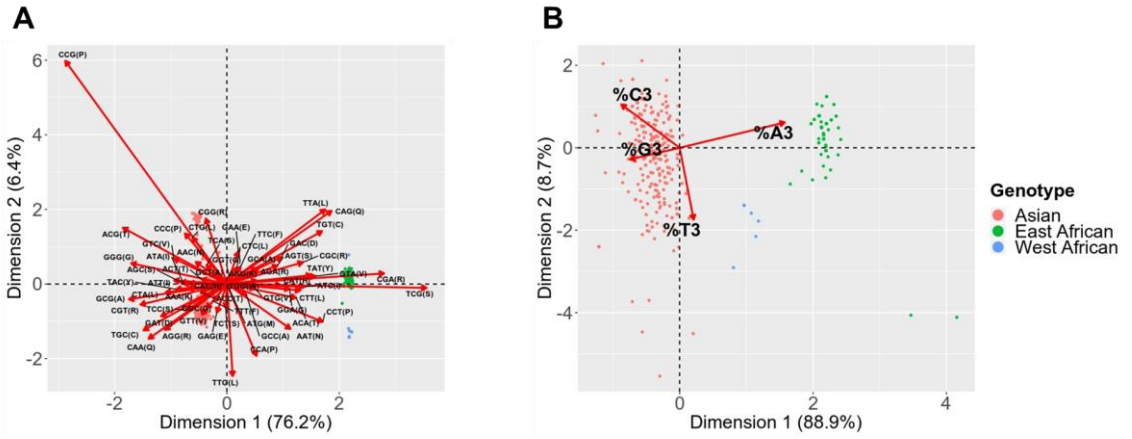

Figure 4. Correspondence analysis plot of Zika virus genotypes (N = 494). The clustering analysis was done using relative synonymous codon usage (RSCU) values (A) and third nucleotides of codons (B). While most RSCU values show little difference between the genotypes, CGA and TCG separate the African genotypes from the Asian genotype. Codons that end with adenine (%A3) and thymine (%T3) are also more prevalent in these genotypes, while codons that end with cytosine (%C3) and guanine (%G3) are more frequent in Asian genotypes. In figure A, dimension 1 explains 76.2 percent and dimension 2 6.4 percent of the variability. In figure B, dimension 1 contributes to 88.9 percent and dimension 2 to 8.7 percent of the variance.

The CA based on nucleotides at the third codon position (%N3) showed similar results (figure 4B) as with RSCUs. Each genotype formed clear and separate clusters. The division between the Asian and African genotype clusters was mostly due to the %G3 and %C3 content in the Asian genotype. The separation of the African genotypes was caused similarly by the %A3 content in the East African genotype and by the %T3 content in the West African genotype. The %N3 in table 3 showed that each genotype had different nucleotide compositions and that each genotype in figure 4B clustered based on their relative %N3.

Table 3. The average third nucleotide composition of codons for each genotype of Zika virus. The percentage values show that there is a bias to use codons that end with guanine (%G3). Because of this and the relatively high cytosine frequency (%C3), the GC3 content of Zika viruses is high compared to the total content of adenine (%A3) and thymine (%T3). The values also show that the variation between the genotypes is minor.

Genotype	%A3	%T3	%C3	%G3	GC3 content
Asian	0.252	0.196	0.256	0.293	0.549
East African	0.263	0.198	0.250	0.288	0.539
West African	0.258	0.198	0.250	0.294	0.544

3.1.2. Dendrograms

In the dendrogram (figure 5A), the three ZKV genotypes formed separate monophyletic clades with the East and West African genotypes being more similar to each other than to the Asian genotype. While viruses from North and South America, and Europe apparently belonged to the Asian genotype, these did not form monophyletic clades. This may indicate that multiple lineages or strains of ZKVs had arrived in America separately. The

dendrogram also revealed that some North American viruses differentiated from South American viruses and vice versa, which might suggest that ZKVs had been spreading between both Americas. Cases of ZKV in Europe could be traced back to American viruses. Most of Oceanic ZKVs formed a clade in the dendrogram, indicating that most viruses in this region originated from a single strain or lineage belonging to the Asian genotype. The dendrogram also shows that ZKVs belonging to the Asian genotype were isolated only from humans and mosquitoes, whereas the African genotypes were mostly isolated from monkeys and mosquitoes.

The unrooted dendrogram (figure 5B) showed similarities between different regional variants of ZKVs with branch lengths. Viruses from Oceania, the Americas and Europe had shorter branch lengths, i.e. their codon usage was most similar to viruses belonging to the Asian genotype. West African, East African and Asian viruses had branch lengths long enough to be clearly distinguishable from each other. The closest relative of ZKVs, Spondweni virus (SPOV), was used in both trees as the outgroup (Haddow and Woodall 2016; Haddow et al. 2016).

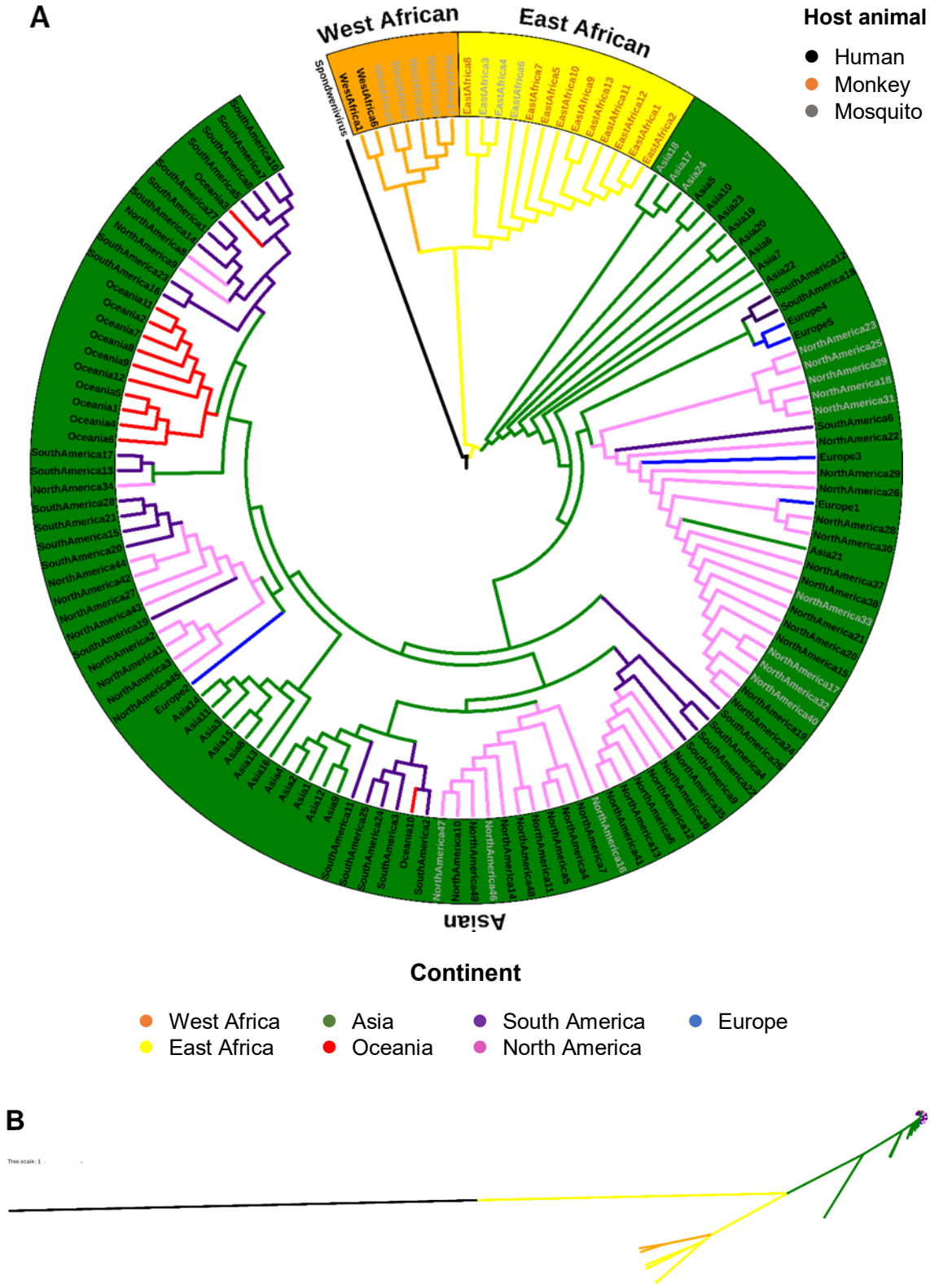

Figure 5. UPGMA dendrograms of Zika viruses based on codon usage values (RSCU) (N = 138). The dendrogram is in two forms: circular (A) and unrooted with proportioned branch lengths (B). Each genotype forms a monophyletic clade and the dendrogram in figure A shows that viruses from Oceania, the Americas and Europe are paraphyletic. Figure A also shows that ZKVs belonging to the Asian genotype were mostly isolated from humans and mosquitoes, whereas the viruses of the African genotypes were mainly from monkeys and mosquitoes. The three genotypes are colored and annotated similarly in both figures. The color of the fonts in figure A shows the host from which the virus was sequenced. In both figures, the color of the branches shows the continent where the virus was found. The dendrogram was calculated with Pearson correlations in DendroUPGMA and rooted with Spondweni virus as the outgroup. The figures are from the poster of the Zika virus pilot study (appendix 2 figure 1).

Another dendrogram was constructed using 44 viral genomes to investigate alternative patterns of ZKV genotypes (figure 6). With the leaves collapsed, the Asian and African genotypes formed two separate monophyletic clades (bootstrap value of 100). The West African genotype also formed a clade (bootstrap value of 99). The least supported clade was the East African genotype (bootstrap values of 26). This result may suggest, that the ZKV started spreading towards Asia and differentiated before the two African genotypes were formed.

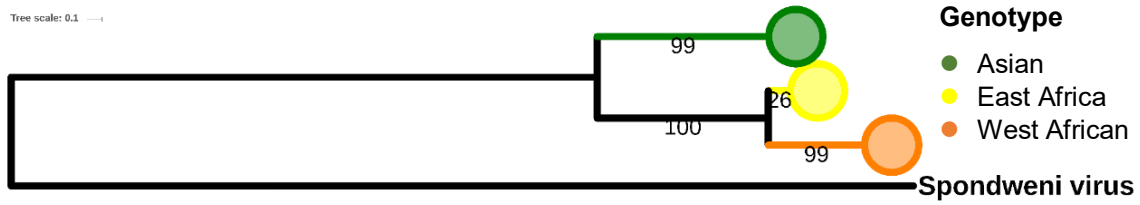

Figure 6. Dendrogram of relative codon usage (RSCU) values of Asian, East and West African genotypes (N = 44). The tree shows (with leaves collapsed) that the African and Asian genotypes form separate clades with bootstrap values of 100 and 99 respectively. This suggests that the Asian genotype differentiated before the two African genotypes were formed. Viruses that were not from the Asian continent were excluded from the Asian genotype. The tree was calculated with Pearson correlations in DendroUPGMA with 100 bootstraps. The branch lengths are proportioned to tree scale. Original figure in the poster of the Zika virus pilot study (appendix 2 figure 1).

3.1.3. Pair-wise distances

The results of the RMSD (table 4) based on average RSCU (appendix 2 table 3) showed the amount of similarity between the ZKVs from different continents. These values revealed, that viruses from Oceania, the Americas and Europe were more similar with Asian viruses than those in Africa. American ZKVs were also slightly more similar to Oceanic viruses (RMSD = 0.012) than to viruses in continental Asia (RMSD = 0.018), which suggests that the virus may have spread through Oceania to Americas. European cases of ZKVs most likely originated from North America (RMSD = 0.008). African ZKVs had low similarity to Asian ones (RMSD > 0.111), thus indicating that they differentiated earlier than the viruses from other locations. East African and West African ZKVs were more similar with each other (RMSD = 0.076), which was in agreement with their common origin. Overall, these results showed that ZKVs in the Western Hemisphere were unlikely to have originated directly from the African continent.

Table 4. Pairwise distance of relative synonymous codon usage (RSCU) values calculated with root-mean-square deviations (RMSD) of Zika viruses from different geographic locations. The values show high similarity in codon usages between Zika viruses sequenced from Americas, Europe and Asia, indicating that these share a common origin. The values also show low similarity between the African genotypes and the Asian genotype, which suggests that the latter genotype evolved from a common African ancestor before the East and West African genotypes differentiated from each other.

	Asia	Oceania	Americas	N-America	S-America	Europe	E-Africa	W-Africa
Asia		0.018	0.018	0.018	0.018	0.017	0.111	0.131
Oceania			0.012	0.013	0.011	0.015	0.113	0.135
Americas				0.003	0.005	0.008	0.114	0.136
N-America					0.008	0.008	0.114	0.135
S-America						0.011	0.115	0.137
Europe							0.115	0.136
E-Africa								0.076
W-Africa								

The RSCU data was also analyzed with a correlation coefficient (figure 7). The results showed very strong correlation ($R^2 > 0.998$) between viruses from continental Asia and Oceania (figure 7A), Americas and Oceania (figure 7B), Americas and Asia (figure 7C), and America and Europe (figure 7D). These findings further indicate that ZKVs from Oceania, the Americas and Europe most likely originated from Asia.

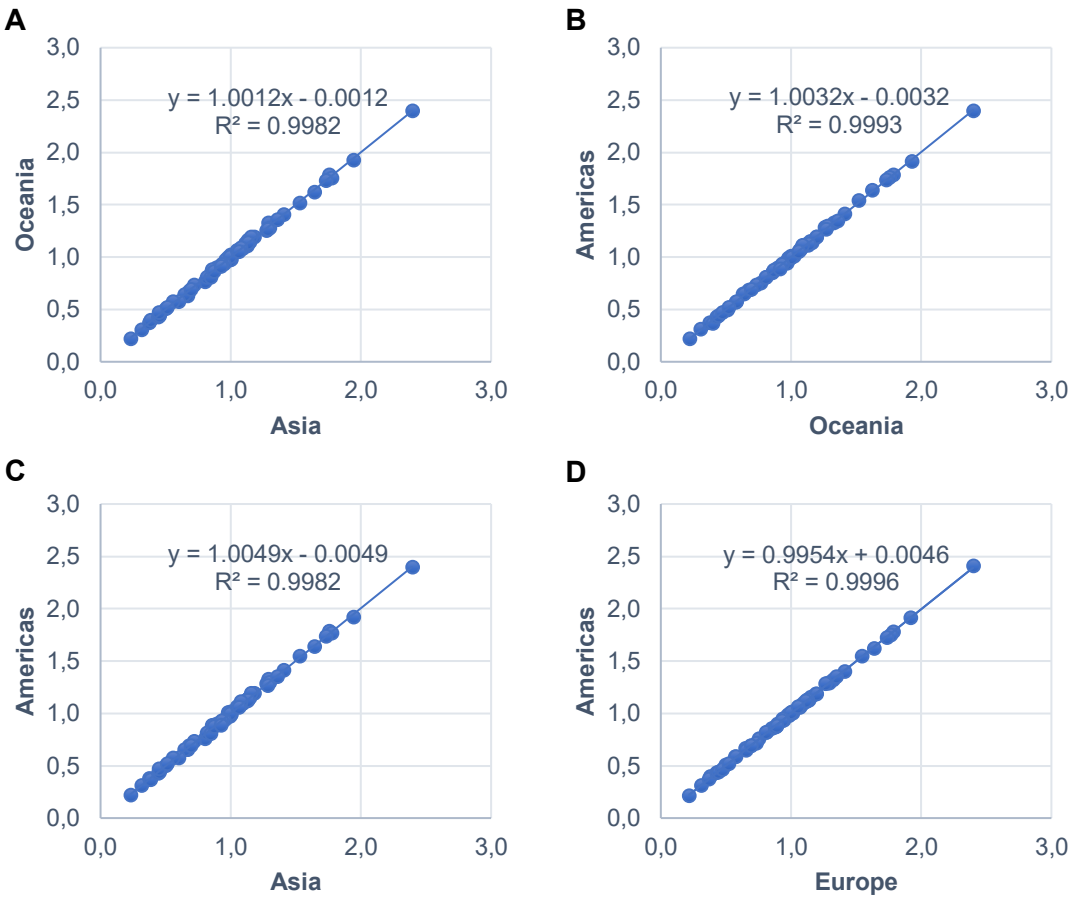

Figure 7. Different determination coefficients between relative synonymous codon usages (RSCU) of Zika viruses from Asia, America, Oceania and Europe. The figures show that the codon usages of American, Oceanic and European viruses are nearly identical to Asian viruses ($R^2 > 0.998$), indicating to a shared and recent evolutionary origin.

3.2. Genomic composition

The methodology described and tested with ZKVs in the previous chapter was applied to DENVs, JEVs and WNVs separately and then with all four major MBFVs together. The RSCU based clustering method proved to be capable of differentiating the subgroups of viruses (genotypes, lineages or serotypes). Based on the provided results, it could be extrapolated that the proposed RSCU based methodology could be used to identify known and possibly novel subgroups of MBFVs. Overall, RSCU was more reliable than %N3 to distinguish intraspecific groups. However, at a species level, the differences were negligible between these two parameters.

3.2.1. Dengue virus

The results of the CA showed that both codon usage (figure 8A) and nucleotides at the third codon positions (figure 8B) were able to differentiate serotypes of DENV. The results were in agreement with the pilot study, as the distances of individual clusters were wider with codon usage than with %N3. The distances between the RSCU clusters did not reflect evolutionary relationships, i.e. were not in agreement with phylogenetic tree constructed with FastTree from AA sequences (figure 9) or literature (Twiddy et al. 2003; Grard et al. 2010). However, %N3 values did match the phylogeny. The clustering patterns of each serotype also differed between the RSCU and %N3 analyses.

The clustering of DENV serotypes based on RSCU values (figure 8A) did not mirror the actual evolutionary relationships (figure 9). Additionally, in figure 8A, within serotypes 1 and 2 were smaller clusters, which indicated that there may have been more subgroups within identified serotypes, which could not be seen in with %N3 values (figure 8B).

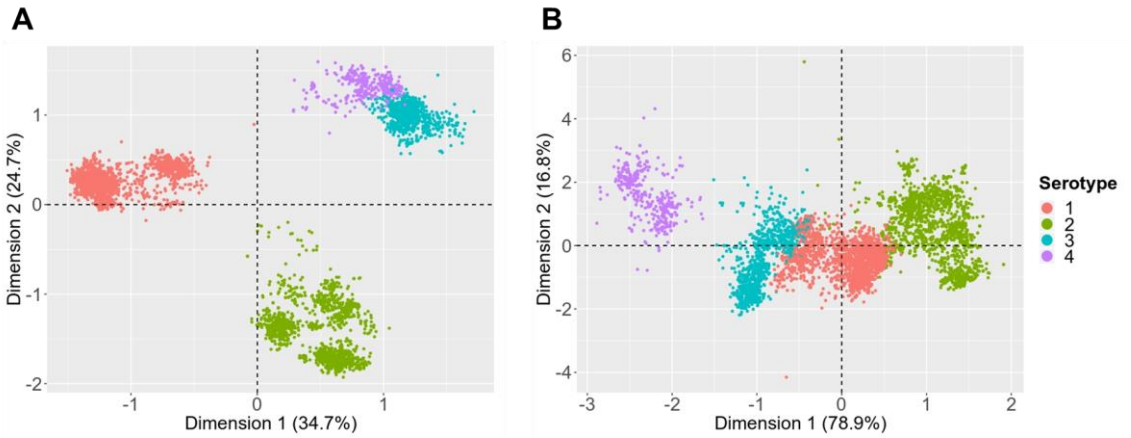

Figure 8. Correspondence analysis plots of Dengue virus genomes (N = 4,865). Plots were computed with relative synonymous codon usages (A) and third nucleotide contents (B). Both figures show that different Dengue virus serotypes cluster together based on either codon usage or nucleotide composition, although the distances between the serotype clusters are higher when using codon usage values. In figure A, dimension 1 explains 34.7 percent and dimension 2 contributes to 24.7 percent of the variation, while in figure B, dimension 1 affects 78.9 percent and dimension 2 explains 16.8 percent of the variation.

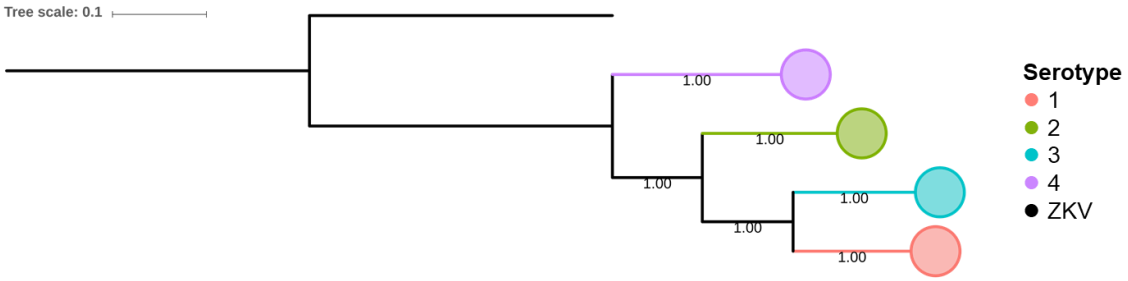

Figure 9. Phylogenetic tree of Dengue viruses (N = 4,865). Each serotype forms a clear and distinct monophyletic clade. The tree was built from protein sequences with 100 bootstraps. Zika virus was used as the outgroup, and the branch lengths are proportioned based on the Tree scale. Clades with multiple sequences are marked as circles.

3.2.2. Japanese encephalitis virus

The CAs based on RSCU (figure 10A) and %N3 (figure 10B) separated JEV genotypes (I–V). Distances between the genotype clusters showed the level of similarity in codon usage and nucleotide composition between these genotypes. The results of genotypes II, IV and V were less conclusive because of the low number of sequences available at NCBI. Although the number of the CDSs for these recently established genotypes was limited, they clustered far enough from the larger genotype I and III, that it could be inferred that genotypes IV and V may have been separate genotypes. Genotype II clustered together with genotype III instead of genotype I, which contradicted the phylogenetic tree (based on a multiple sequence alignment of proteins and FastTree) of JEVs (figure 11). Nonetheless, the results based on the CA were mostly in agreement with the phylogeny.

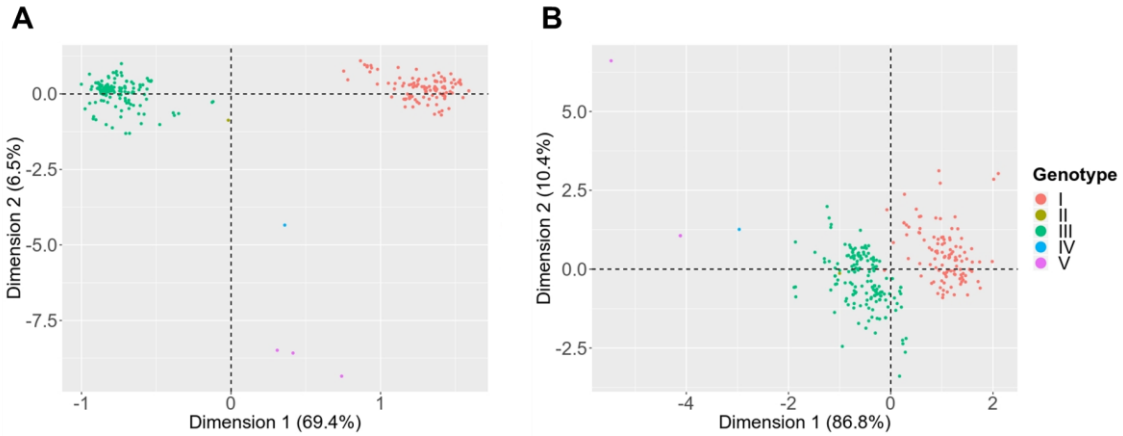

Figure 10. Correspondence analysis of Japanese encephalitis virus genomes (N = 297). The plots were computed with relative codon usage (RSCU) values (A) and with third nucleotides of each codon (B). Each genotype forms distinct clusters, which are farther from each other with RSCUs than with third nucleotide content. In plot A, dimension 1 explains 69.4 percent and dimension 2 6.5 percent of the observed variation. In plot B, dimension 1 explains 86.8 percent and dimension 2 10.4 percent of the variation.

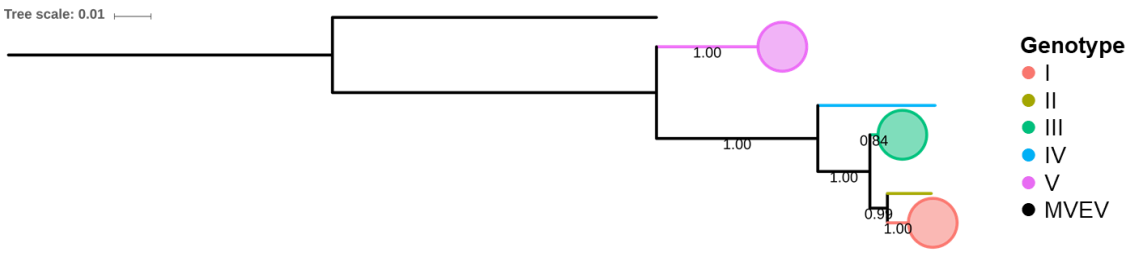

Figure 11. Phylogenetic tree of Japanese encephalitis viruses (N = 297). The tree is built from amino acid sequences with 100 bootstraps. Murray valley encephalitis virus (MVEV) was used as an outgroup and the lengths of the branches match the Tree scale. Branches with circles at the end have multiple sequences collapsed.

3.2.3. West Nile virus

The WNV sequences formed three larger clusters based on their RSCU and %N3 values (figure 12). The main clusters were formed by lineages 1A, 1B and 2. There were, however, differences in clustering depending whether RSCU or %N3 was used. Based on the codon usage, lineages 4 and 5 clustered close to lineage 2, and lineages 7, 8 and 9/4c were spread out without forming a clear cluster. The second analysis based on %N3s showed more similarity between lineages 1A and 1B. Lineage 5 also shared more similarities with 1A than 2. Lineage 4 clustered farther from the other lineages with %N3 values. Lineage 7 consistently clustered near lineage 2. The recently proposed lineage 9/4c also had different clustering patterns depending on the analysis. This lineage was equidistant to any genotype based on the RSCU but was closer to lineage 1A based on %N3. The results based on the CAs were in agreement with those from the phylogenetic tree of WNV AA sequence alignments and FastTree (figure 13).

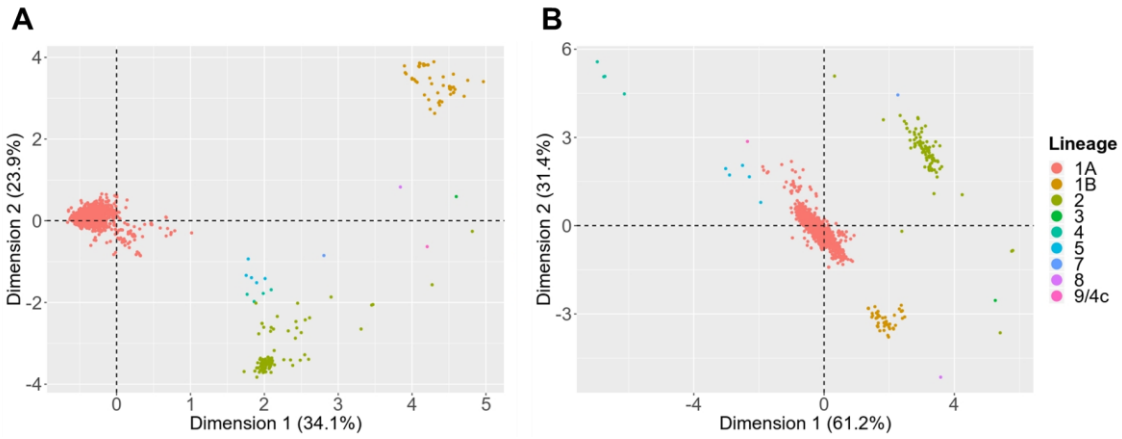

Figure 12. Correspondence analysis plot of West Nile viruses (N = 1,619). The figures were computed with codon usage values (A) and compositions of nucleotides in third codon position (B). The distances between the lineage clusters are on average longer with codon usages than with nucleotide content. In figure A, dimension 1 explains 34.1 percent of the variation and dimension 2 23.9 percent. In figure B, dimension 1 attributes to 61.2 percent of variance and dimension 2 to 31.4 percent.

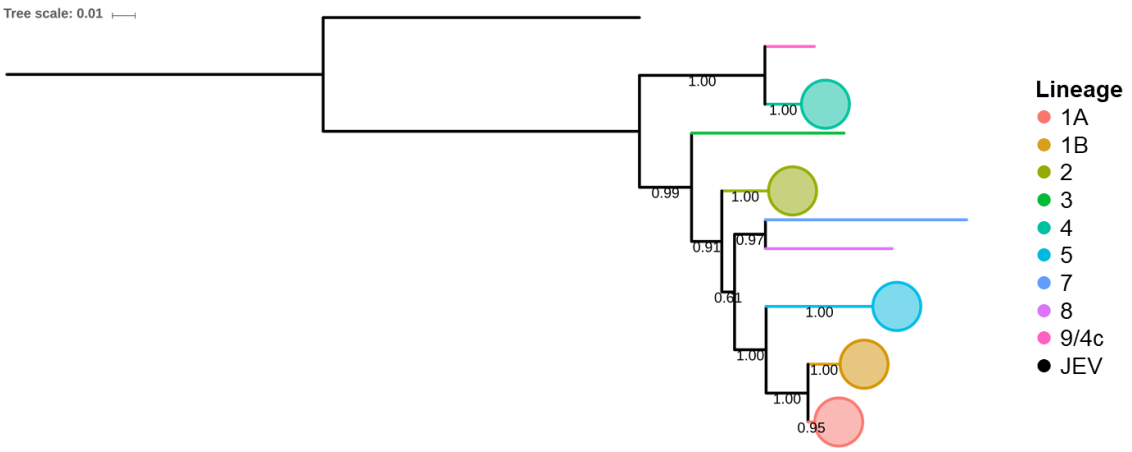

Figure 13. Phylogenetic tree of West Nile virus genomes (N = 1,619). Protein sequences of these viruses were used to construct the tree. The tree was computed with 100 bootstraps and the reference sequence of Japanese encephalitis virus was used as an outgroup to root the tree. The branch lengths are proportioned to the Tree scale. Branches that have circles at the end contain multiple virus sequences and lines with none have only one.

3.2.4. Zika virus

The results showed that RSCU values distinguished the East African, West African and Asian genotypes from each other. Based on the RSCU analysis, East and West African genotypes clustered together and apart from the Asian genotype (figure 14A), which was in agreement with the “Africa/Asia” hypothesis (de Bernardi Schneider et al. 2016) of the evolutionary history of ZKV genotypes. Additionally, RSCU values split the Asian genotype to two very distinct subgroups and formed on average tighter clusters compared to third nucleotide compositions. While %N3 differentiated the three genotypes, it was not to the same degree as RSCU (figure 14B). It also did not match the phylogenetic tree constructed from AA sequence alignments with FastTree (figure 15), as the West African genotype cluster was located in the middle of the Asian and East African genotypes, which made it hard to infer evolutionary relationships between them.

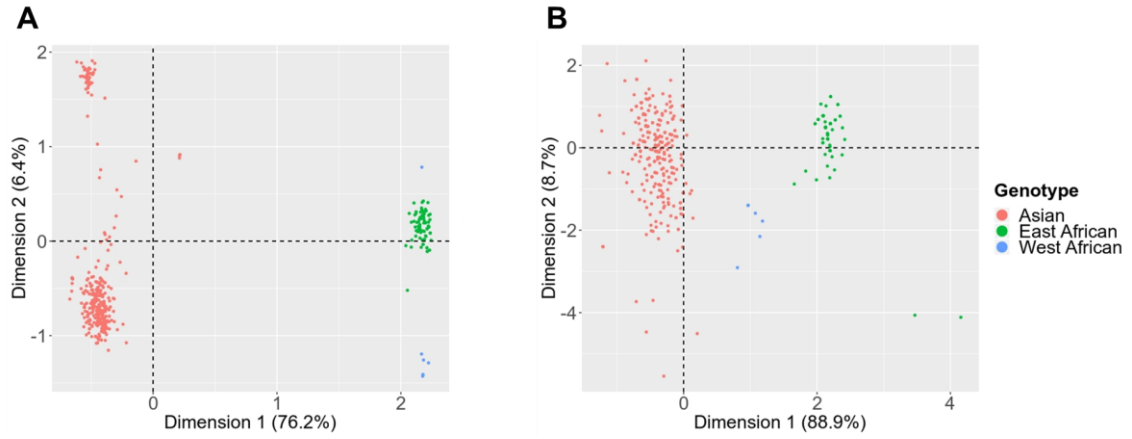

Figure 14. Correspondence analysis of RSCU and %N3 of Zika viruses (N = 494). Analyses conducted with relative synonymous codon usages (RSCUs) (A) and compositions based on the third nucleotide of each codon (%N3) (B) show, that each Zika virus genotype form separate clusters. The values separate the African genotypes from the Asian one. RSCU distinguishes the genotypes more compared to %N3 values and forms tighter clusters. RSCU additionally splits the Asian genotype into two smaller clusters. In figure A, dimension 1 explains 76.2 percent of the variation, while dimension 2 explains 6.4 percent. In figure B, dimension 1 contributes to 88.9 percent of the variation and dimension 2 to 8.7 percent.

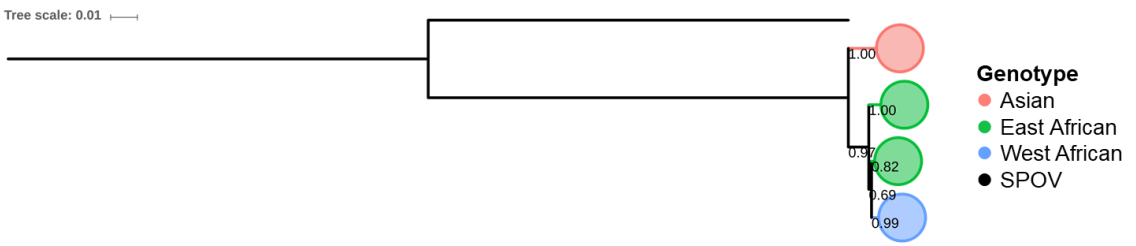

Figure 15. Phylogenetic tree of Zika virus genomes (N = 494). The tree was computed using amino acid sequences based on complete coding sequences and 100 bootstraps. Spondweni virus (SPOV) was used as an outgroup to root the tree. The length of branches is proportioned according to the Tree scale.

3.2.5. Major mosquito-borne flaviviruses

The CA of RSCU and %N3 of four major MBFVs showed that interspecies clustering patterns did not change much between RSCU and %N3 values (figure 16), although there were differences in clustering patterns. The CA based on the RSCU showed that JEVs and WNVs clustered together, whereas in the CA based on the %N3 ZKVs were closer to WNVs than to JEVs. The major flaviviruses tended to separate based on RSCU and %N3 into two groups: one with DENVs and another with JEVs, WNVs and ZKVs. This was in accordance to the phylogenetic tree computed from protein sequences and constructed with FastTree (figure 17).

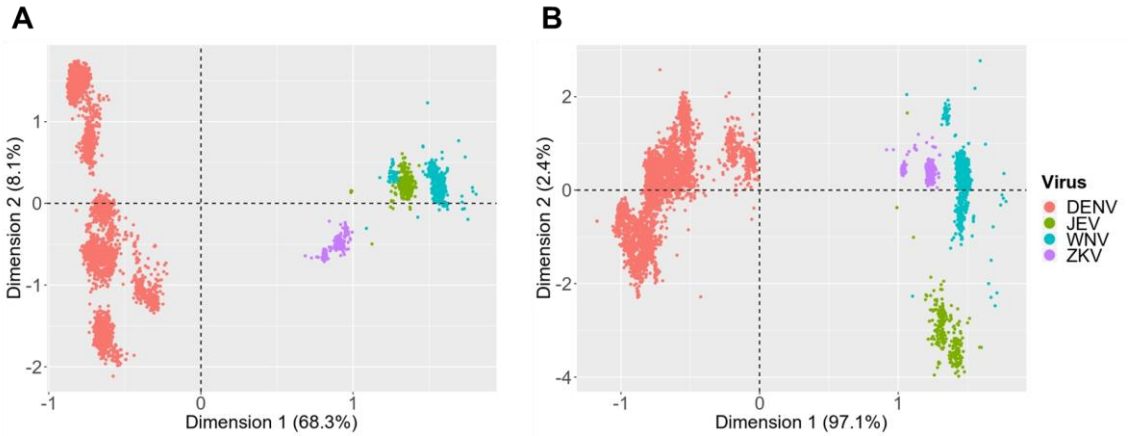

Figure 16. Correspondence analysis plots of the four major mosquito-borne flavivirus genomes (N = 7,274). The viruses analyzed were Dengue virus (DENV), Japanese encephalitis virus (JEV), West Nile virus (WNV) and Zika virus (ZKV). Figure A was done using relative synonymous codon usage (RSCU) values and figure B with third nucleotide contents of each codon (%N3). In both plots, the distances between DENVs and the other viruses is almost equal, which means that on a species level, RSCU and %N3 contents are similar as differentiating factors. In figure A, 68.3 percent of variation is explained by dimension 1, and 8.1 percent with dimension 2. In figure B, dimension 1 contributes to 97.1 percent of variance and dimension 2 to 2.4 percent.

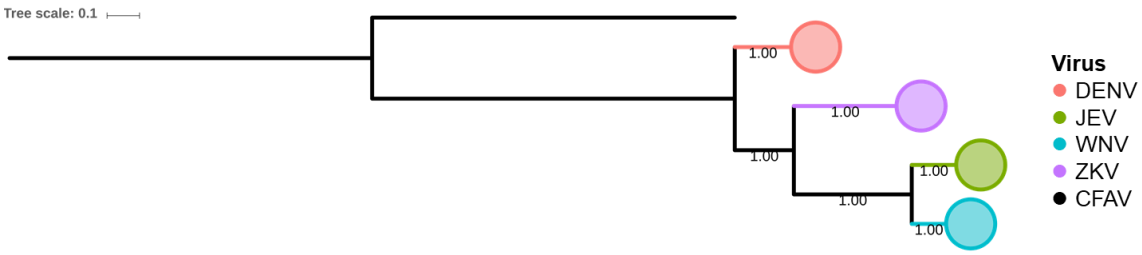

Figure 17. Phylogenetic tree of four major mosquito-borne flaviviruses (N = 7,274). The tree was built from the coding sequences of Dengue virus (DENV), Japanese encephalitis virus (JEV), West Nile virus (WNV) and Zika virus (ZKV) and computed with 100 bootstraps. The branch lengths are scaled to match the Tree scale. The tree was then rooted to the outgroup virus, Cell fusing agent virus (CFAV). If a virus had multiple sequences, they were collapsed and marked with circles.

3.3. Optimal host identification

The optimal hosts of flaviviruses were identified by comparing the codon usages between viruses and putative hosts. The comparison was done by analyzing nCAI values (CAI_h divided by CAI_s) with heatmaps and CAs. The CA results in figure 18 revealed inter- and intraspecific patterns. On average, most mammalian hosts and *Aedes* mosquitoes were optimal for flaviviruses (nCAI 0.95–1.05), while birds and other arthropods (*Culex* and *Anopheles* mosquitoes, and ticks) were suboptimal (nCAI <0.95 or >1.05) (figure 19). This stemmed from overadaptation to avian hosts and from underadaptation to the other arthropods. While these codon usage patterns seemed uniform across flaviviruses, the correlations between nCAI values showed that each subgroup had a specific optimal host organisms.

3.3.1. Optimal hosts

Figure 18 shows that different subgroups of flaviviruses formed separate clusters based on nCAI values, i.e. the value of adaptation to putative hosts. The further a virus was from the center of the CA plot (origin), the more similar, and therefore more optimized, codon usage it had towards an organism. MBFVs (use mosquito-vectors) and UVFVs (have an unidentified vector) clustered towards vertebrates, IOFVs (infect only insects) towards *Aedes* mosquitoes and TBFVs (use tick-vectors) towards deer tick (*Ixodes scapularis*). The dhIOFVs (infect both insects and vertebrates) formed a cluster between *Aedes* mosquitoes and vertebrates. The clusters formed roughly two groups based on their optimization for a host type: vertebrates and mosquitoes, which are further illustrated in supplementary material (appendix 3 figures 2 and 3). The vertebrate cluster included MBFVs, TBFVs, UVFVs and dhIOFVs, while the mosquito cluster included only IOFVs. None of the subgroups clustered towards *Anopheles* or *Culex* mosquitos, suggesting that none of them were optimal hosts. The subgroup classification of each flavivirus (in addition to their recorded hosts and vectors) was based on literature (appendix 1 table 1).

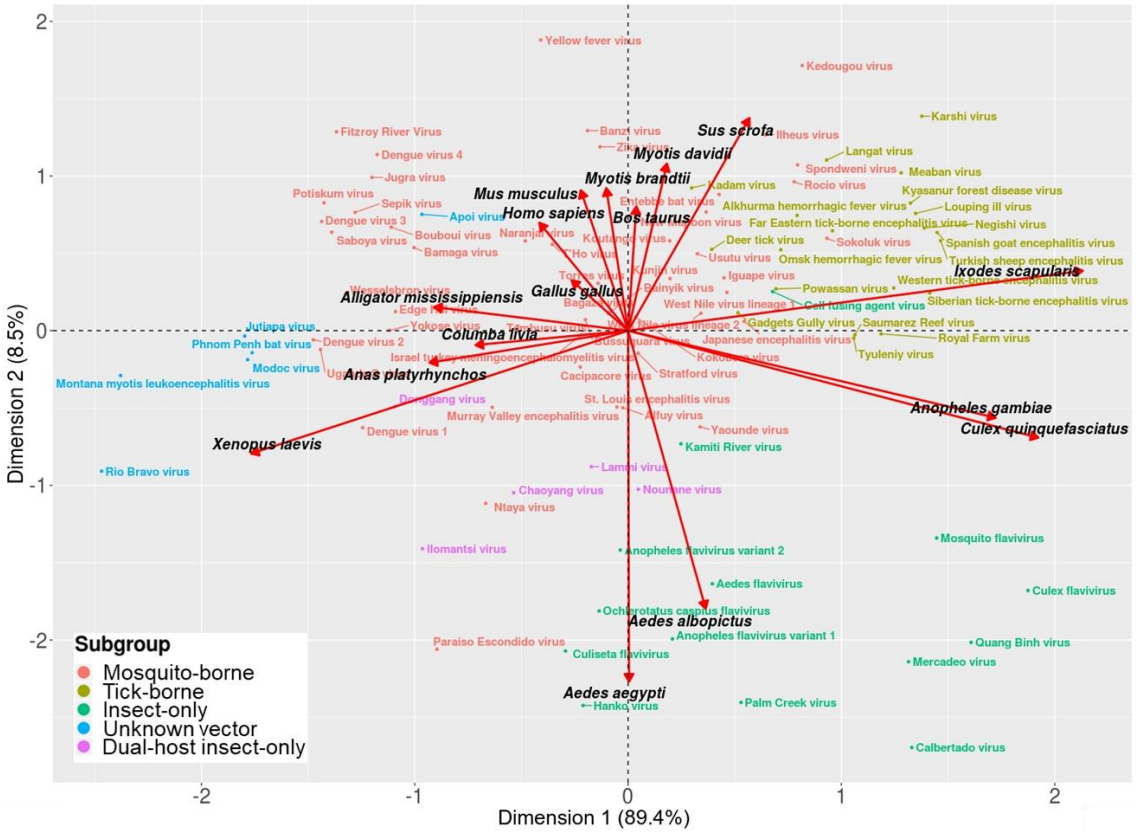

Figure 18. Correspondence analysis of normalized Codon Adaptation Index (nCAI) values of flaviviruses (N = 94). The plot shows that nCAI can differentiate multiple subgroups of flaviviruses based on the degree of codon usage optimization for host organisms. Mosquito-borne flaviviruses are generally optimized to vertebrate hosts, while tick-borne flaviviruses are towards ticks and insect-only flaviviruses are optimized for mosquitoes. Dual-host insect-only flaviviruses show optimization for both *Aedes* mosquitoes and vertebrates, and unknown vector flaviviruses are optimized for vertebrates. Dimension 1 explains 89.4 percent of the variation and dimension 2 explains 8.5 percent.

3.3.2. Trends in codon usage optimization

In the heatmap, with flavivirus subgroups arranged according to phylogenetic trees (appendix 3 figure 1) and with hosts sorted based on taxonomical classification, the nCAI values showed adaptation patterns that were common across all viruses (figure 19). The results for arthropod and vertebrate hosts were as follows:

Arthropod hosts: The analysis suggested that *Aedes* mosquitoes were the optimal host for all flaviviruses, whereas *Culex* mosquitoes, *Anopheles gambieae* and deer tick showed low adaptation to all viruses, indicating that they were unlikely hosts. The deer tick was, however, optimal for TBFVs, which use ticks as carrying vectors.

Vertebrate hosts: Most of the mammals were within the range of likely hosts, which included bats (*Myotis brandtii* and *My. davidii*), house mouse (*Mus musculus*), cattle (*Bos taurus*) and humans (*Homo sapiens*). The wild boar (*Sus scrofa*) seemed to be on average the least optimal host for all flaviviruses but notice that the CUT available was incomplete (see methods section). Birds (*Columba livia*, *Gallus gallus* and *Anas platyrhynchos*), amphibians (*Xenopus laevis*) and reptiles (*Alligator mississippiensis*) showed on average overadaptation, except in the case of IOFVs. This may suggest that they are suboptimal hosts.

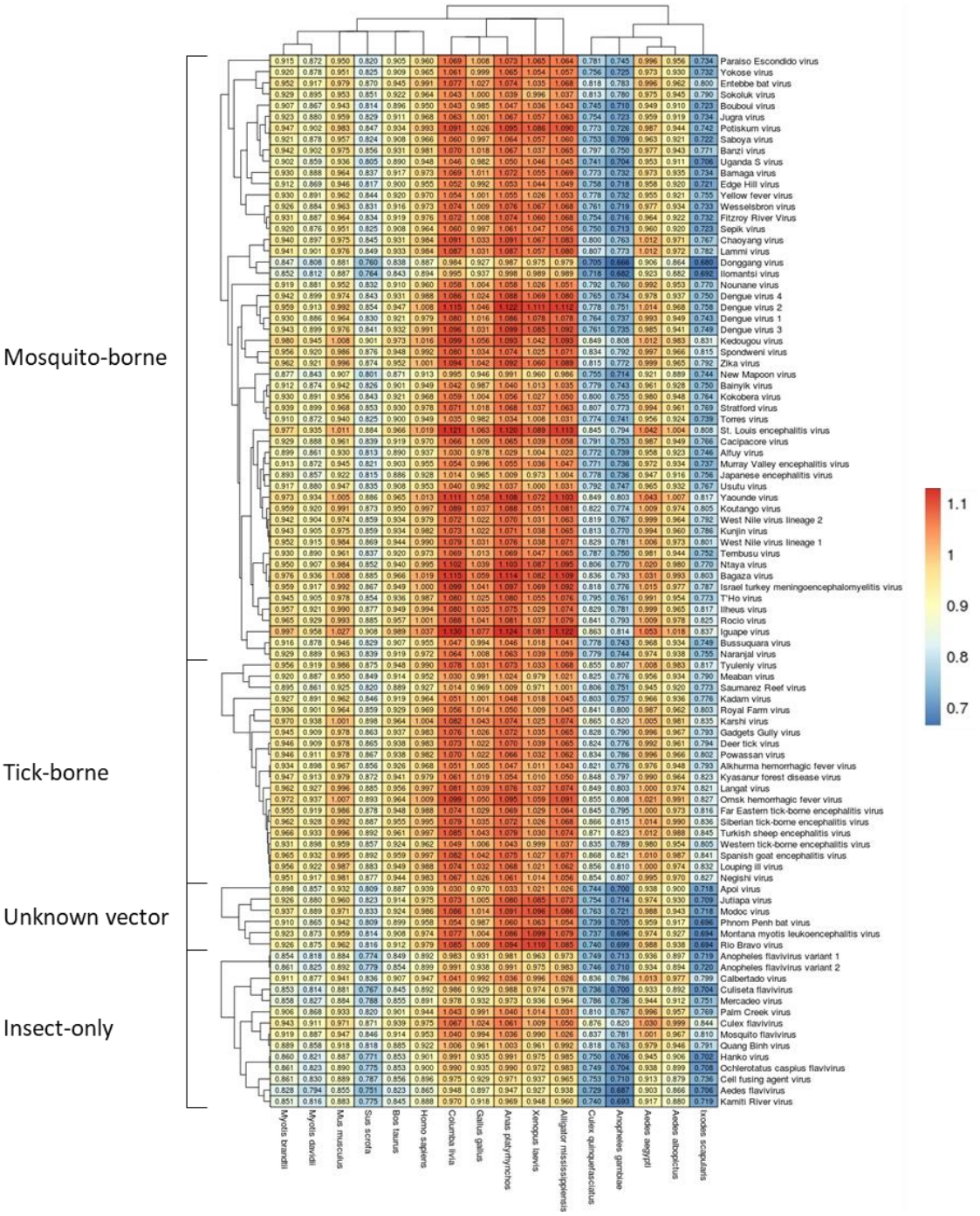

Figure 19. Normalized Codon Adaptation Index (nCAI) heatmap of flavivirus (genus *Flavivirus*) subgroups (N = 94). The nCAI values show overall overoptimization (nCAI > 1.05) for avians, reptiles and amphibians, and underoptimization for *Culex* and *Anopheles* mosquitoes (nCAI < 0.95). Optimal hosts tend to be mammals and *Aedes* mosquitoes (nCAI 0.95–1.05). The columns are sorted according to taxonomic classification of hosts, and the rows are in accordance to the phylogeny of flaviviruses.

4. DISCUSSION

4.1. Examination of the Zika virus pilot study

The tested methodology clustered each of the ZKV genotypes accurately into monophyletic clades within the dendrogram. This tree cannot be strictly interpreted as phylogenetic but produced clades that reflect actual phylogenetic trees based on proteins (appendix 2 figure 2). These results suggest, that all ZKVs outside the African continent belong to the Asian genotype, which aligns with contemporary knowledge (Yokoyama and Starmer 2017; May and Relich 2016; Gong et al. 2017; Faria et al. 2016). Additionally, the computed dendrograms support the proposed “Africa/Asia” hypothesis (de Bernardi Schneider et al. 2016) as the correct alternative to explain the differentiation of ZKV genotypes.

The analysis based on RSCU values revealed codon usage patterns within ZKVs. While the variations in RSCU values between genotypes were minor, the values were distinct enough to differentiate the genotypes in CA. Additionally, the results showed that American and European ZKVs belong to the Asian genotype. These findings are in agreement with other studies from the scientific literature (Cristina et al. 2016; Singh and Tyagi 2017; Butt et al. 2016; Wang et al. 2016; van Hemert and Berkhout 2016), although each study had minor variations between individual values probably due to the significantly lower number of viral sequences used. Overall, the results suggest that methods based on codon usage patterns are able to reveal variations between ZKV genotypes. Thus, it can be assumed that they can be utilized to analyze other flaviviruses. This methodology might represent a step forward to a reliable and faster characterization of flaviviruses.

4.1.1. Validity of proposed methodology

The implemented methods in the pilot study were able to create dendrograms that reflected the actual phylogenetic relationships among ZKVs on a genotypic level (appendix 2 figure 2). Sequences clustered into three monophyletic clades with high bootstrap values (figure 6). Each of these clades had sequences of the same genotype as reported in NCBI and Virus Pathogen Resource (appendix 2 table 2), thus proving that the viruses could form clusters based on their codon usages. The methods used in the pilot study were able to accurately discern all three ZKV genotypes despite the minute differences in codon usages (appendix 2 table 3) and third nucleotide contents. Based on these results, it is within reason to infer, that this method could be applied to identify other viruses that may have similarly small genomic variations. The applicability of the proposed methodology was further tested in this thesis, and its results will be discussed later.

4.1.2. Differences found among Zika viruses

Each of the ZKV genotype had slight deviations in codon usage that differed from the species average, but none of these was significantly different (appendix 2 table 3). Variations of the codon usage were however distinct enough to differentiate the three major genotypes into separate monophyletic clades of the UPGMA dendrogram (figure 5).

One of the aims of the pilot study was to investigate, whether the American ZKVs were able to be identified based on their codon usage, in other words, whether they would cluster into a monophyletic group. According to the results, while there were very subtle differences in RSCUs, the viruses from America did not have any identifiable usage pattern (appendix 2 table 3). This conclusion can be also inferred from the bootstrap values computed for the Asian genotype (figure 20). Most of the clades that were produced from the clustering of American viruses did not have enough bootstrap support to justify a new genotype.

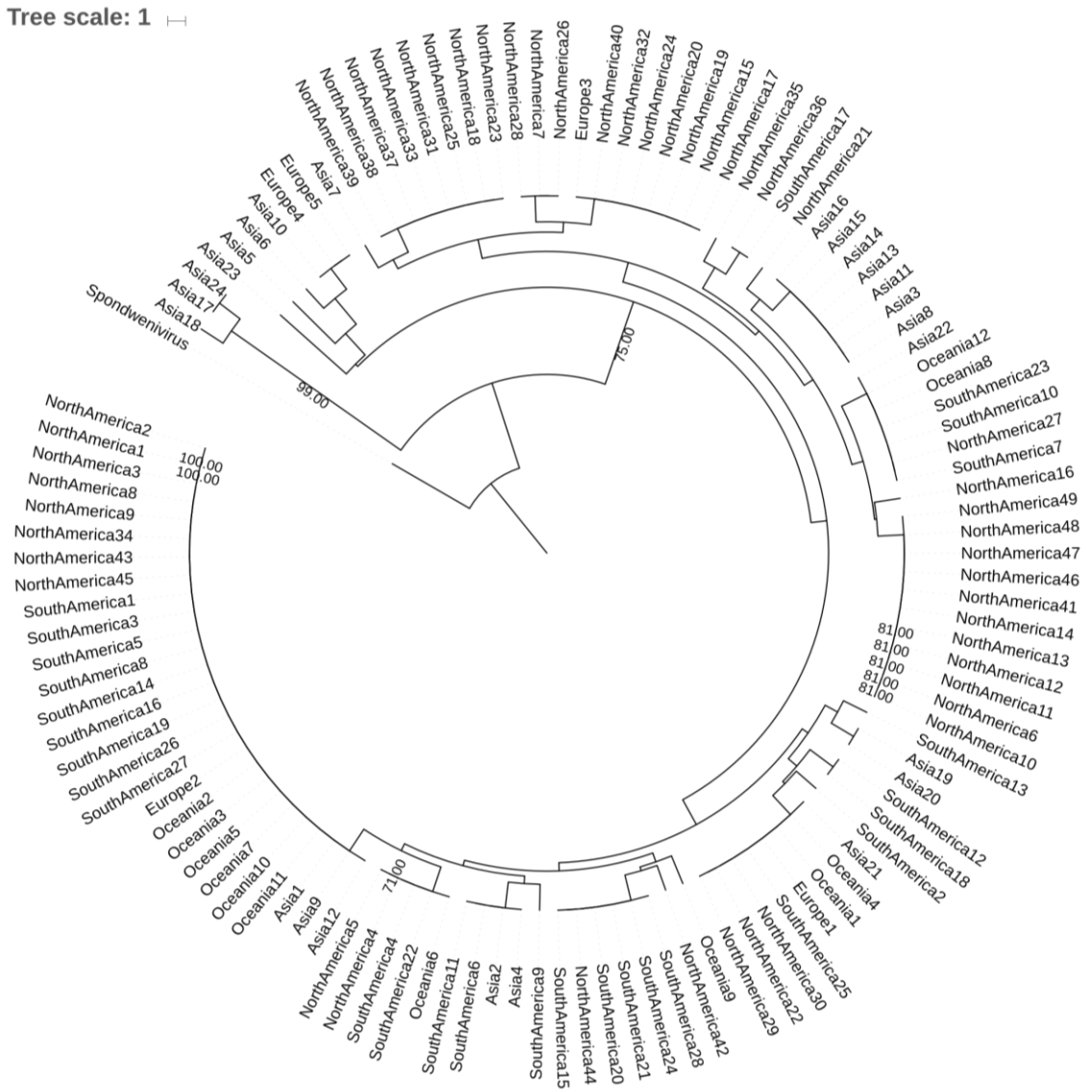

Figure 20. Dendrogram of Asian genotype of Zika viruses. Tree was computed from relative codon usages (RSCU) with Pearson correlations and 100 bootstraps using DendroUPGMA. Only bootstrap values above 75 are shown, and the length of branches is proportioned to Tree scale. The tree shows no significant clustering of sequences belonging to the Asian genotype with the exceptions of two North American subclades and one Asian subgroup. Tree was rooted with the reference sequence of Spondweni virus. Sequence information can be found in appendix 2 table 2.

The analysis based on %N3 also provided a way to differentiate the genotypes. Quantitative nucleotide compositions showed small variations between genotypes (table 3), which is in agreement with the results seen based on RSCU values, but with lower resolution. Nonetheless, the analysis of %N3 showed enough genotype specificity to be implemented in a genotype determination process along with RSCU values.

4.1.3. Origin of Zika virus

The origin of ZKVs is well documented (Kindhauser et al. 2016; Posen et al. 2016), but how the three genotypes differentiated from each other is still under debate (Gong et al. 2016). Two alternative hypotheses have been proposed (figure 21): the “Africa/Asia” hypothesis assumes that the Asian genotype and the ancestral African genotype separated before the African genotype split into the current East and West African genotypes, whereas the “Out of Africa” hypothesis suggests that the African genotypes are lineages that were formed when the virus was spreading towards Asia (de Bernardi Schneider et al. 2016). Different phylogenetic methods give results that support either of these hypotheses, probably due to the low amount of genetic differences between the genotypes. Usually methods based on maximum parsimony support the first hypothesis and maximum-likelihood in turn supports the latter hypothesis (de Bernardi Schneider et al. 2016).

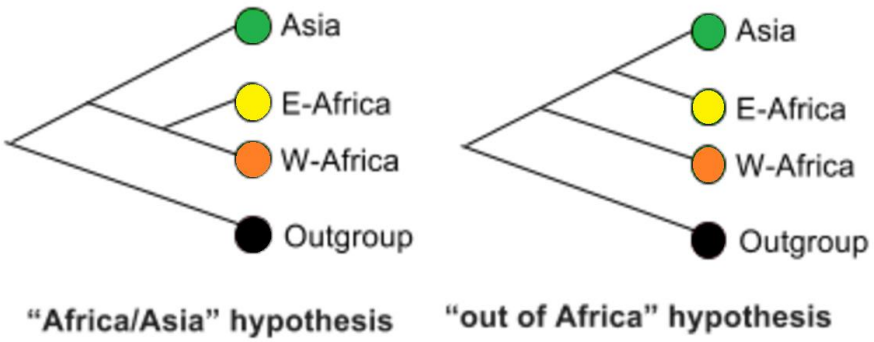

Figure 21. Proposed hypotheses for the differentiation of Zika virus genotypes. According to “Africa/Asia” hypothesis, the Asian genotype differentiated from African genotypes early, while according to “Out of Africa” hypothesis, the Asian genotype split off the African genotypes later when it started to spread towards Asia (de Bernardi Schneider et al. 2016). In the study, Spondweni virus was used as an outgroup. Original figure can be found in appendix 2 figure 1.

The results produced in this pilot study supported the aforementioned “Africa/Asian” hypothesis, which was also supported by a control tree constructed from AA sequences (appendix 2 figure 2). The calculated bootstrap values lock the African clades together, which consequently means that the Asian genotype separated from the African genotypes before East and West African genotypes were formed. Similar evolutionary patterns were also found in other studies (van Hemert and Berkhout 2016).

Additional examination of the dendrogram revealed, that the Asian genotype forms four subclades based on their codon usage; two North American (bootstrap values of 81 and 100) and two Asians (bootstrap values of 99 and 75) (figure 20). When the countries and collection dates of the sequences in North American subclades are compared, both provide evidence that one route of spreading from South America was through the Caribbean Islands (mainly Puerto Rico and Martinique) to Central America. This is also revealed in the North American subclade (bootstrap values of 100) that shows that a ZKV

from the island of Martinique infected a human host in Chiapas, Mexico, although the infection could have been produced in Mexico or carried from Martinique. Also, the analysis of the Asian subclade suggests that a group of Malaysian ZKVs did not continue its spread, unlike the other strains that spread globally.

4.2. Genomic composition of major mosquito-borne flaviviruses

The results of the RSCU and %N3 analyses of DENVs, JEVs and WNVs are consistent with the results established with ZKV pilot study. In each case, both values could differentiate the subgroups within a species. However, the results of certain subgroups of JEV and WNV (with a limited number of available complete genomes) are inconclusive. In general, the RSCU analysis was more accurate to separate subgroups of flaviviruses, and the clustering patterns were usually in agreement with the actual phylogeny. For example, RSCUs were able to identify country specific subpopulations within the Asian genotype of ZKVs, which were Malaysia and Singapore. It is likely however, that these cases originated from a single strain in their respective countries (Haddow et al. 2012; Singapore Zika Study Group 2017). The RSCU analysis is able to distinguish interspecies subgroups, but the amount intraspecies variability does not change significantly. Similar studies, based on a few number of representative sequences, have been done on DENVs (Ma et al. 2013; Lara-Ramírez et al. 2014), JEVs (Singh et al. 2016), WNVs (Moratorio et al. 2013) and ZKVs (Butt et al. 2016; Cristina et al. 2016). Each study support the results presented here.

The different subgroups of flaviviruses clustered according their specific codon usage preferences. These preferences are known to be affected by the codon usage biases of their respective host organisms (Bahir et al. 2009). RSCUs are more informative than %N3 values due to the larger number of variables in consideration, i.e. RSCU values are based on 59 codon, while %N3 has only 4 nucleotide variables. Surprisingly, even though the number of variables in %N3 was low, it was still able to separate flavivirus subgroups. This suggests that through high mutation rates affecting viral genomes, there was a high amount of genomic diversity found even within a species of virus. These findings indicate that RSCUs may be used as a factor to identify optimal hosts.

4.3. Optimal host identification of flaviviruses

4.3.1. Aspects of codon usage optimization and nCAI

An important quality of the CAI is the capability to measure the adaptation between the codon usages of a virus to its host organism. This assumption is based on the idea that the more similar the codon usage of a virus and the host are, the more adapted or specialized the virus is to an organism. The changes in codon usage in viruses to match the host are akin to the process of gene amelioration (Lawrence and Ochman 1997), in which

the nucleotide composition of a foreign genetic material will over time change to match the composition of its recipient.

There are several reasons to explain why viruses optimize their codon usage to imitate their hosts. It may occur to increase translation rate (Tuller et al. 2010) and thus replication in host cells (Karlin et al. 1990), to match the specific transfer RNA (tRNA) composition of the host to maximize protein synthesis rates (Ikemura 1981; Zhou et al. 1999; Michely et al. 2013), or to avoid translational errors due to an abundance of rare codons (Kane 1995). All of these factors increase the survivability and infectiveness of viruses. This pattern has been observed several times between viruses and their hosts (Kunisawa et al. 1998; Bahir et al. 2009; Lobo et al. 2009).

However, this trend is not followed by certain viruses that use an opposite strategy. It has been observed that certain viruses maintain a degree of deoptimization in codon usage to maximize their range of hosts (Butt et al. 2016), to avoid triggering and prolong the activation of the immune system of the host (Mossadegh et al. 2004; Zhou et al. 1999; Cid-Arregui et al. 2003; Karlin et al. 1990; Zhou et al. 2012), and to ensure proper protein folding (Zhou et al. 2012). This, however, lowers the viral replication due to increased attenuation (Gao et al. 2003; Zhou et al. 2012).

While the reasons for underoptimization based on nCAI can be explained by low similarity in codon usages, the reasons for overoptimization are not as clearly defined. One explanation for this could be the over expression of certain genes in the host, which inadvertently affects the codon composition of the infecting virus by highly increasing its bias toward certain synonymous codons. The advantages with reproducing via highly expressed host genes may impose a translational selection pressure (Sharp et al. 1993) towards extremely optimized codon usages, thus promoting specialization towards certain hosts and, furthermore, narrow the range of possible host organisms. The link between optimal codon composition, and high gene expression and reproduction rates have been observed in rapidly multiplying bacteria (Rocha 2004). However, a high optimization in codon usage has a detrimental effect, for it increases probability of inducing a stronger immune response from the host (Zhao and Chen 2011; Ramakrishna et al. 2004) and raises the likelihood of improper protein folding (Aragonès et al. 2010). Larger DNA viruses may have a lesser need for perfect codon usage optimization compared to smaller DNA viruses, because they usually have genes to encode proteins that inhibit immune responses of the host (Shackelton et al. 2006).

The optimal range of hosts can be better inferred from the CA plot (figure 18). The results suggest that MBFVs generally have a life cycle, in which a virus uses vertebrates as hosts. TBFVs have a similar host preference, i.e. vertebrates, but additionally have certain adaptation to spread through ticks. The nCAI-CA analysis is able to classify that ticks

are involved in the life cycle of TBFVs without the addition of any prior knowledge. As expected, the analysis suggests that IOFVs primarily infect insects. The dhIOFVs can infect both insect and vertebrate hosts and this is reflected in their location between the centroid clouds of the mosquito and vertebrate clusters (appendix 3 figure 2). The CA plots show that codon usage values are discriminative on both intraspecies and interspecies level.

To investigate, whether the proposed host identification methodology could distinguish different subgroups within a viral species, the results of nCAI were analyzed and plotted via CA (appendix 3 figure 5). The viruses chosen for this were DENV, JEV, WNV and ZKV mostly due to their current and historic prevalence as highly virulent pathogens capable of causing lethal diseases in humans and because of the high number of available CDSs. The nCAI could be used to differentiate subgroups of flaviviruses similarly to RSCU and %N3. The applicability and accuracy of the proposed algorithm was additionally tested by analyzing the genomic variability within a species. The intraspecies variability was analyzed with all available CDSs of DENVs, JEVs, WNVs and ZKVs. The analysis of the nCAI values with k-means clustering confirmed that reference sequences used for DENVs, JEVs, WNVs and ZKVs were representative of the entire pool of sequences (appendix 3 figure 4).

4.3.2. Estimated hosts and their accuracy

The analysis of the CA from nCAI values suggest that MBFVs should have a reproductive cycle which included a mosquito vectors from genus *Aedes*, and a primary mammalian host. The low adaptation of MBFVs according to nCAI to *Culex* and *Anopheles* mosquitoes indicate they are unlikely used as host, but this does not mean, that *Culex* and *Anopheles* mosquitoes are completely incapable of occasionally carrying *Aedes* specific MBFVs (Dodson and Rasgon 2017; Amraoui et al. 2016; Althouse et al. 2015). *Culex* and *Anopheles* mosquitoes were suboptimal hosts across all flaviviruses, but the range of optimal hosts varied depending on the viral species.

Several of the estimated hosts were in agreement with previous studies and observations (appendix 1 table 1). For example, according to the CA-nCAI analysis, WNV lineages 1 and 2 could potentially infect ticks, because they are located near the TBFV cluster. This is supported by the results obtained by Lawrie et al. (2004). It has been also observed that the primary host of Aedes flavivirus (an IOFV) is the (Asian) tiger mosquito (*Aedes albopictus*) (Hoshino et al. 2009), but the virus has also been found in *Culex* mosquitoes (Grisenti et al. 2015). The CA-nCAI results indicate that Aedes flavivirus is clearly more optimized to *Aedes* mosquitoes, thus the presence of sequences from this virus in *Culex* should be accidental. Finally, the analysis was also able to identify a potential host type for Usutu virus (an MBFV). The virus favors bird hosts (Meister et al. 2008), yet according

to nCAI, Usutu virus also has optimized codon usage towards mammals. This finding is supported by two cases of Usutu virus causing infections in humans (Pecorari et al. 2009; Cavrini et al. 2009).

The algorithm also provided interesting results for the paraphyletic subgroup of dhIOFVs. The optimization of most dhIOFVs for both vertebrate and mosquito hosts could be observed with nCAI and this was in agreement with the description of the subgroup. The reason for this dual optimization in dhIOFVs may be due to either having unknown vertebrate hosts or having lost the ability to infect them (Blitvich and Firth 2015).

There were, however, some inconsistencies with the results despite the relatively high level of accuracy. The algorithm grouped all vertebrates close together, causing difficulties to distinguish the exact optimal host. For example, Yokose virus (an MBFV) is optimized for birds based on nCAI, although it mainly infects bats. The CA-nCAI analysis also suggests that JEVs are optimized for ticks, although observations in nature prove otherwise (appendix 1 table 1). Lastly, Alfuy virus (an MBFV) has been found mostly in *Culex* mosquitoes (Colmant et al. 2017) and yet the results show more optimization towards *Aedes* mosquitoes. Overall, the identified optimal hosts mostly matched the ones mentioned in literature.

5. CONCLUSIONS

The aim of this thesis was to provide a computationally quantitative method to estimate the likelihood of a flavivirus to infect a host organism. This method was based on the analysis of the codon usage and a normalized version of the Codon Adaptation Index (nCAI) to quantify the adaptation of a virus to the host. The accuracy of this methodology was first tested on a smaller pool of Zika viruses (ZKVs). This pilot study supported the assumptions that different subgroups within a virus could be differentiated by their distinct relative use of synonymous codons (RSCU) and nucleotide composition at the third codon position (%N3). Remarkably both factors reflected the phylogeny of the ZKV.

This genomic characterization methodology based on the codon usage was further tested with ZKVs and three additional major flaviviruses, Dengue viruses (DENVs), Japanese encephalitis viruses (JEVs) and West Nile viruses (WNVs). The results of these analyses were similar to the pilot study, thus differences among flaviviruses can be traced through RSCU and %N3 patterns.

A normalized version of the classical Codon Adaptation Index (nCAI) was next used to identify codon optimization levels between flaviviruses and the range of potential host organisms. The results of this analysis indicate that flaviviruses are optimized to two major groups: vertebrate and mosquito hosts. Moreover, tick-borne flaviviruses (TBFVs) formed a minor subcluster within vertebrates. These findings were supported by current literature.

The proposed nCAI based optimal host identification algorithm presented in this thesis provides a simple tool to identify putative hosts, and thus, establish a plausible range of risk hosts to be monitored. This methodology could be used efficiently as part of a surveillance system, because it does not need prior knowledge other than coding sequences from query viruses and potential hosts. A scientific article describing the host identification algorithm is under preparation (appendix 4).

6. FUTURE IMPROVEMENTS

The results of this master's thesis show that the nCAI and codon usage based methodologies are reliable to estimate the host of flaviviruses. In future research, this methodology could be improved by combining additional, and independent, parameters based on interactions between (1) membrane proteins and receptors of viruses and hosts, (2) cellular translation machinery and viral RNA, and (3) host immune responses and virus control.

The current work assessed as many potential hosts as possible to be representative, but CUTs are not currently available for multiple hard and soft tick species, most notably castor bean tick (*Ixodes Ricinus*) and taiga tick (*Ix. persulcatus*), and sand fly genera (*Psathyromyia*, *Phlebotomus* and *Sergentomyia*) from which Paraiso Escondido virus and Saboya virus have been sequenced (Alkan et al. 2015; Fontenille et al. 1994; Ba et al. 1999).

With additional minor improvements, the proposed methodology could be used to predict putative hosts in other types of viruses, or to identify the hosts of novel viruses during future outbreaks.

7. ACKNOWLEDGMENTS

I sincerely thank Pere Puigbò, PhD for his guidance and support as a mentor and supervisor during all phases of this master's thesis, and for the tools and materials he provided.

I would like to thank Santi Garcia-Vallvé, PhD for providing the DendroUPGMA tree construction utility and his contributions in developing and publishing the host identification methodology.

I would like to give thanks to Tiina Henttinen, PhD for giving advice and ensuring that the styling, formatting and other aspects of this thesis were in accordance to the rulings of the Department of Biology.

I would like to thank IT Center for Science (CSC) for providing the servers and generous resources for performing computationally heavy analyses.

I would also thank the University of Turku and the Department of Biology for providing working spaces and equipment for this thesis.

Finally, I thank my family, relatives and friends for their continued "support".

8. REFERENCES

Agüero M, Fernández-Pinero J, Buitrago D, Sánchez A, Elizalde M, San Miguel E, Villalba R, Llorente F, Jiménez-Clavero MA. 2011. Bagaza virus in partridges and pheasants, Spain, 2010. *Emerging Infect Dis* **17**: 1498–1501.

Alkan C, Zapata S, Bichaud L, Moureau G, Lemey P, Firth AE, Gritsun TS, Gould EA, de Lamballerie X, Depaquit J, et al. 2015. Ecuador Paraiso Escondido Virus, a New Flavivirus Isolated from New World Sand Flies in Ecuador, Is the First Representative of a Novel Clade in the Genus Flavivirus. *J Virol* **89**: 11773–11785.

Althouse BM, Hanley KA, Diallo M, Sall AA, Ba Y, Faye O, Diallo D, Watts DM, Weaver SC, Cummings DAT. 2015. Impact of climate and mosquito vector abundance on sylvatic arbovirus circulation dynamics in Senegal. *Am J Trop Med Hyg* **92**: 88–97.

Amberg SM, Rice CM. 1999. Mutagenesis of the NS2B-NS3-mediated cleavage site in the flavivirus capsid protein demonstrates a requirement for coordinated processing. *J Virol* **73**: 8083–8094.

Amraoui F, Atyame-Nten C, Vega-Rúa A, Lourenço-de-Oliveira R, Vazeille M, Failloux AB. 2016. Culex mosquitoes are experimentally unable to transmit Zika virus. *Euro Surveill* **21**.

Ando K, Kuratsuka K, Arima S, Hironaka N, Honda Y, Ishii K. 1952. Studies on the Viruses isolated during Epidemic of Japanese B Encephalitis in 948 in Tokyo Area. *Kitasato Archives of Experimental Medicine*.

Aragonès L, Guix S, Ribes E, Bosch A, Pintó RM. 2010. Fine-tuning translation kinetics selection as the driving force of codon usage bias in the hepatitis A virus capsid. *PLoS Pathog* **6**: e1000797.

Arnal A, Gómez-Díaz E, Cerdà-Cuéllar M, Lecollinet S, Pearce-Duvet J, Busquets N, García-Bocanegra I, Pagès N, Vittecoq M, Hammouda A, et al. 2014. Circulation of a Meaban-like virus in yellow-legged gulls and seabird ticks in the western Mediterranean basin. *PLoS One* **9**: e89601.

Asghar N, Lindblom P, Melik W, Lindqvist R, Haglund M, Forsberg P, Överby AK, Andreassen Å, Lindgren P-E, Johansson M. 2014. Tick-borne encephalitis virus sequenced directly from questing and blood-feeding ticks reveals quasispecies variance. *PLoS One* **9**: e103264.

Ba Y, Trouillet J, Thonnon J, Fontenille D. 1999. [Phlebotomus of Senegal: survey of the fauna in the region of Kedougou. Isolation of arbovirus]. *Bull Soc Pathol Exot* **92**: 131–135.

Badrane H, Tordo N. 2001. Host switching in Lyssavirus history from the Chiroptera to the Carnivora orders. *J Virol* **75**: 8096–8104.

Bahir I, Fromer M, Prat Y, Linial M. 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* **5**: 311.

Bancroft WH, Scott RM, Snitbhan R, Weaver RE, Gould DJ. 1976. Isolation of Langat virus from *Haemaphysalis papuana* Thorell in Thailand. *Am J Trop Med Hyg* **25**: 500–504.

Barnston AG. 1992. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Wea Forecasting* **7**: 699–709.

Batista WC, Tavares G da SB, Vieira DS, Honda ER, Pereira SS, Tada MS. 2011. Notification of the first isolation of Cacipacore virus in a human in the State of Rondônia, Brazil. *Rev Soc Bras Med Trop* **44**: 528–530.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res* **41**: D36–42.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, et al. 2013. The global distribution and burden of dengue. *Nature* **496**: 504–507.

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

Bolling BG, Eisen L, Moore CG, Blair CD. 2011. Insect-specific flaviviruses from Culex mosquitoes in Colorado, with evidence of vertical transmission. *Am J Trop Med Hyg* **85**: 169–177.

Bondre VP, Jadi RS, Mishra AC, Yergolkar PN, Arankalle VA. 2007. West Nile virus isolates from India: evidence for a distinct genetic lineage. *J Gen Virol* **88**: 875–884.

Bondre VP, Sapkal GN, Yergolkar PN, Fulmali PV, Sankararaman V, Ayachit VM, Mishra AC, Gore MM. 2009. Genetic characterization of Bagaza virus (BAGV) isolated in India and evidence of anti-BAGV antibodies in sera collected from encephalitis patients. *J Gen Virol* **90**: 2644–2649.

Bowen ETW, Simpson DIH, Platt GS, Way HJ, Smith CEG, Ching CY, Casals J. 1970. Arbovirus infections in Sarawak: The isolation of Kunjin virus from mosquitoes of the *Culex pseudovishnui* group. *Annals of Tropical Medicine & Parasitology* **64**: 263–268.

Boyle DB, Dickerman RW, Marshall ID. 1983. Primary viraemia responses of herons to experimental infection with murray valley encephalitis, kunjin and japanese encephalitis viruses. *Aust J Exp Biol Med* **61**: 655–664.

Braverman Y, Boreham PFL, Galun R, Ziv M. 1977. The origin of blood meals of biting midges (Diptera: Ceratopogonidae) and mosquitoes (Diptera: Culicidae) trapped in turkey runs in Israel. *Rhod J Agric Res*.

Braverman Y, Davidson I, Chizov-Ginzburg A, Chastel C. 2003. Detection of Israel Turkey Meningo-encephalitis Virus from Mosquito (Diptera: Culicidae) and *Culicoides* (Diptera: Ceratopogonidae) Species and Its Survival in *Culex pipiens* and *Phlebotomus papatasi* (Diptera: Phlebotomidae). *J Med Entomol* **40**: 518–521.

Butt AM, Nasrullah I, Qamar R, Tong Y. 2016. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg Microbes Infect* **5**: e107.

Campbell GL, Hills SL, Fischer M, Jacobson JA, Hoke CH, Hombach JM, Marfin AA, Solomon T, Tsai TF, Tsu VD, et al. 2011. Estimated global incidence of Japanese encephalitis: a systematic review. *Bull World Health Organ* **89**: 766–74, 774A.

Cavrini F, Gaibani P, Longo G, Pierro AM, Rossini G, Bonilauri P, Gerunda GE, Benedetto FD, Pasetto A, Girardis M, et al. 2009. Usutu virus infection in a patient who underwent orthotropic liver transplantation, Italy, August-September 2009. *Eurosurveillance*.

CDC. 1985. Arbovirus Catalog: Yaounde (YAOV). *Centers for Disease Control and Prevention*. https://wwwn.cdc.gov/arbocat/VirusDetails.aspx?ID=528&SID=7 (Accessed July 23, 2018).

CDC. 2017. West Nile Virus Disease Cases by State. *Centers for Disease Control and Prevention*. https://www.cdc.gov/westnile/statsmaps/preliminarymapsdata2017/disease-cases-state.html (Accessed June 7, 2018).

Chambers TJ, Hahn CS, Galler R, Rice CM. 1990. Flavivirus genome organization, expression, and replication. *Annu Rev Microbiol* **44**: 649–688.

Charrel RN, Zaki AM, Fakeeh M, Yousef AI, de Chesse R, Attoui H, de Lamballerie X. 2005. Low diversity of Alkhurma hemorrhagic fever virus, Saudi Arabia, 1994-1999. *Emerging Infect Dis* **11**: 683–688.

Chen WR, Rico-Hesse R, Tesh RB. 1992. A new genotype of Japanese encephalitis virus from Indonesia. *Am J Trop Med Hyg* **47**: 61–69.

Chen WR, Tesh RB, Rico-Hesse R. 1990. Genetic variation of Japanese encephalitis virus in nature. *J Gen Virol* **71 (Pt 12)**: 2915–2922.

Cid-Arregui A, Juárez V, zur Hausen H. 2003. A synthetic E7 gene of human papillomavirus type 16 that yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization studies. *J Virol* **77**: 4928–4937.

Cleaves GR, Dubin DT. 1979. Methylation status of intracellular dengue type 2 40 S RNA. *Virology* **96**: 159–165.

Codon Usage Database. 2017. Codon Usage Database. http://www.kazusa.or.jp/codon/ (Accessed June 27, 2017).

Coimbra TL, Nassar ES, Nagamori AH, Ferreira IB, Pereira LE, Rocco IM, Ueda-Ito M, Romano NS. 1993. Iguape: a newly recognized flavivirus from São Paulo State, Brazil. *Intervirology* **36**: 144–152.

Colmant AMG, Bielefeldt-Ohmann H, Hobson-Peters J, Suen WW, O'Brien CA, van den Hurk AF, Hall RA. 2016. A newly discovered flavivirus in the yellow fever virus group displays restricted replication in vertebrates. *J Gen Virol* **97**: 1087–1093.

Colmant AMG, Hobson-Peters J, Bielefeldt-Ohmann H, van den Hurk AF, Hall-Mendelin S, Chow WK, Johansen CA, Fros J, Simmonds P, Watterson D, et al. 2017. A New Clade of Insect-Specific Flaviviruses from Australian Anopheles Mosquitoes Displays Species-Specific Host Restriction. *mSphere* **2**.

Cook S, Bennett SN, Holmes EC, De Chesse R, Moureau G, de Lamballerie X. 2006. Isolation of a new strain of the flavivirus cell fusing agent virus in a natural mosquito population from Puerto Rico. *J Gen Virol* **87**: 735–748.

Coz J, Valade M, Cornet M, Robin Y. 1976. [Transovarian transmission of a Flavivirus, the Koutango virus, in Aedes aegypti L]. *C R Acad Sci Hebd Seances Acad Sci D* **283**: 109–110.

Cristina J, Fajardo A, Soñora M, Moratorio G, Musto H. 2016. A detailed comparative analysis of codon usage bias in Zika virus. *Virus Res* **223**: 147–152.

CSC. 2017. CSC – IT Center for Science. https://www.csc.fi/ (Accessed January 26, 2018).

Das SR, Puigbò P, Hensley SE, Hurt DE, Bennink JR, Yewdell JW. 2010. Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. *PLoS Pathog* **6**: e1001211.

Davies FG. 1978. Nairobi sheep disease in Kenya. The isolation of virus from sheep and goats, ticks and possible maintenance hosts. *J Hyg* **81**: 259.

de Bernardi Schneider A, Malone RW, Guo J-T, Homan J, Linchangco G, Witter ZL, Vinesett D, Damodaran L, Janies DA. 2016. Molecular evolution of Zika virus as it crossed the Pacific to the Americas. *Cladistics*.

de Souza Lopes O, Coimbra TL, de Abreu Sacchetta L, Calisher CH. 1978a. Emergence of a new arbovirus disease in Brazil. I. Isolation and characterization of the etiologic agent, Rocio virus. *Am J Epidemiol* **107**: 444–449.

de Souza Lopes O, de Abreu Sacchetta L, Coimbra TL, Pinto GH, Glasser CM. 1978b. Emergence of a new arbovirus disease in Brazil. II. Epidemiologic studies on 1975 epidemic. *Am J Epidemiol* **108**: 394–401.

de Souza Lopes O, de Abreu Sacchetta L, Francy DB, Jakob WL, Calisher CH. 1981. Emergence of a new arbovirus disease in Brazil. III. Isolation of Rocio virus from Psorophora Ferox (Humboldt, 1819). *Am J Epidemiol* **113**: 122–125.

Dick GWA, Haddow AJ. 1952. Uganda S virus. A hitherto unrecorded virus isolated from mosquitoes in Ugnada. (I). Isolation and pathogenicity. *Trans R Soc Trop Med Hyg* **46**: 600–618.

Dick GWA, Kitchen SF, Haddow AJ. 1952. Zika virus. I. Isolations and serological specificity. *Trans R Soc Trop Med Hyg* **46**: 509–520.

Dobson AP, Carper ER. 1996. Infectious diseases and human population history. *Bioscience* **46**: 115–126.

Dodson BL, Rasgon JL. 2017. Vector competence of Anopheles and Culex mosquitoes for Zika virus. *PeerJ* **5**: e3096.

Doherty RL, Carley JG, Kay BH, Filippich C, Marks EN, Frazier CL. 1979. Isolation of virus strains from mosquitoes collected in Queensland, 1972-1976. *Aust J Exp Biol Med Sci* **57**: 509–520.

Doherty RL, Standfast HA, Domrow R, Wetters EJ, Whitehead RH, Carley JG. 1971. Studies of the epidemiology of arthropod-borne virus infections at Mitchell River Mission, Cape York Peninsula, North Queensland IV. Arbovirus infections of mosquitoes and mammals, 1967–1969. *Trans R Soc Trop Med Hyg* **65**: 504–513.

Doherty RL, Whitehead RH, Judith Wetters E, Gorman BM. 1968. Studies of the epidemiology of arthropod-borne virus infections at Mitchell River Mission, Cape York Peninsula, North Queensland. *Trans R Soc Trop Med Hyg* **62**: 430–438.

ECDC. 2017. Transmission of West Nile fever, May to November 2017. *European Centre for Disease Prevention and Control*. https://ecdc.europa.eu/en/publications-data/table-transmission-west-nile-fever-may-november-2017-table-cases-2017 (Accessed June 7, 2018).

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Efron B. 1979. Bootstrap Methods: Another Look at the Jackknife. *Ann Statist* **7**: 1–26.

Estrada-Peña A, Jongejan F. 1999. Ticks feeding on humans: a review of records on human-biting Ixodoidea with special reference to pathogen transmission. *Exp Appl Acarol* **23**: 685–715.

Failloux A-B, Vazeille M, Rodhain F. 2002. Geographic genetic variation in populations of the dengue virus vector Aedes aegypti. *J Mol Evol* **55**: 653–663.

Fall G, Diallo M, Loucoubar C, Faye O, Sall AA. 2014. Vector competence of Culex neavei and Culex quinquefasciatus (Diptera: Culicidae) from Senegal for lineages 1, 2, Koutango and a putative new lineage of West Nile virus. *Am J Trop Med Hyg* **90**: 747–754.

Farfan-Ale JA, Loroño-Pino MA, Garcia-Rejon JE, Hovav E, Powers AM, Lin M, Dorman KS, Platt KB, Bartholomay LC, Soto V, et al. 2009. Detection of RNA from a novel West Nile-like virus and high prevalence of an insect-specific flavivirus in mosquitoes in the Yucatan Peninsula of Mexico. *Am J Trop Med Hyg* **80**: 85–95.

Faria NR, Azevedo R do S da S, Kraemer MUG, Souza R, Cunha MS, Hill SC, Thézé J, Bonsall MB, Bowden TA, Rissanen I, et al. 2016. Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**: 345–349.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res* **40**: D136–43.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Felsenstein J. 2004. Finding the best tree by heuristic search. In *Inferring Phylogenies*, pp. 38–44.

Ferreira DD, Cook S, Lopes Â, de Matos AP, Esteves A, Abecasis A, de Almeida APG, Piedade J, Parreira R. 2013. Characterization of an insect-specific flavivirus (OCFVPT) co-isolated from Ochlerotatus caspius collected in southern Portugal along with a putative new Negev-like virus. *Virus Genes* **47**: 532–545.

Fields BN, Knipe DM, Howley PM, Griffin DE. 2001. Chapter 33 Flaviviruses. In *Fields Virology*, Vol. 2 of, p. 891, Lippincott Williams & Wilkins, Philadelphia.

Firth AE, Atkins JF. 2009. A conserved predicted pseudoknot in the NS2A-encoding sequence of West Nile and Japanese encephalitis flaviviruses suggests NS1' may derive from ribosomal frameshifting. *Virol J* **6**: 14.

Fontenille D, Traore-Lamizana M, Diallo M, Thonnon J, Digoutte JP, Zeller HG. 1998. New vectors of Rift Valley fever in West Africa. *Emerging Infect Dis* **4**: 289–293.

Fontenille D, Traore-Lamizana M, Trouillet J, Leclerc A, Mondo M, Ba Y, Digoutte JP, Zeller HG. 1994. First isolations of arboviruses from phlebotomine sand flies in West Africa. *Am J Trop Med Hyg* **50**: 570–574.

Frost MJ, Zhang J, Edmonds JH, Prow NA, Gu X, Davis R, Hornitzky C, Arzey KE, Finlaison D, Hick P, et al. 2012. Characterization of virulent West Nile virus Kunjin strain, Australia, 2011. *Emerging Infect Dis* **18**: 792–800.

Furuse Y, Suzuki A, Oshitani H. 2010. Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries. *Virol J* **7**: 52.

Gao F, Li Y, Decker JM, Peyerl FW, Bibollet-Ruche F, Rodenburg CM, Chen Y, Shaw DR, Allen S, Musonda R, et al. 2003. Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro expression and immune responses in DNA-vaccinated mice. *AIDS Res Hum Retroviruses* **19**: 817–823.

Garcia-Vallvé S, Puigbò P. 2002. DendroUPGMA: Dendrogram construction using the UPGMA algorithm. *DendroUPGMA: A dendrogram construction utility*. http://genomes.urv.cat/UPGMA/ (Accessed January 9, 2017).

GenBank. 2017. GenBank. https://www.ncbi.nlm.nih.gov/genbank/ (Accessed July 27, 2017).

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**: 1369–1372.

Githeko AK, Lindsay SW, Confalonieri UE, Patz JA. 2000. Climate change and vector-borne diseases: a regional analysis. *Bull World Health Organ*.

Gomes G, Causey OR. 1959. Bussuquara, A New Arthropod-Borne Virus. *Exp Biol Med* **101**: 275–279.

Gong Z, Gao Y, Han G-Z. 2016. Zika virus: two or three lineages? *Trends Microbiol* **24**: 521–522.

Gong Z, Xu X, Han G-Z. 2017. The diversification of Zika virus: Are there two distinct lineages? *Genome Biol Evol* **9**: 2940–2945.

Gordon Smith CE. 1956. A Virus Resembling Russian Spring–Summer Encephalitis Virus from an Ixodid Tick in Malaya. *Nature* **178**: 581–582.

Gould EA, de Lamballerie X, de A. Zanotto PM, Holmes EC. 2001. Evolution, epidemiology, and dispersal of flaviviruses revealed by molecular phylogenies. Vol. 57 of *Advances in virus research*, pp. 71–103, Elsevier.

Gould EA, de Lamballerie X, Zanotto PM d. A, Holmes EC. 2003. Origins, evolution, and vector/host coadaptations within the Genus Flavivirus. Vol. 59 of *Advances in virus research*, pp. 277–314, Elsevier.

Grard G, Moureau G, Charrel RN, Holmes EC, Gould EA, de Lamballerie X. 2010. Genomics and evolution of Aedes-borne flaviviruses. *J Gen Virol* **91**: 87–94.

Grisenti M, Vázquez A, Herrero L, Cuevas L, Perez-Pastrana E, Arnoldi D, Rosà R, Capelli G, Tenorio A, Sánchez-Seco MP, et al. 2015. Wide detection of Aedes flavivirus in north-eastern Italy--a European hotspot of emerging mosquito-borne diseases. *J Gen Virol* **96**: 420–430.

Haddow AD, Nasar F, Guzman H, Ponlawat A, Jarman RG, Tesh RB, Weaver SC. 2016. Genetic Characterization of Spondweni and Zika Viruses and Susceptibility of Geographically Distinct Strains of Aedes aegypti, Aedes albopictus and Culex quinquefasciatus (Diptera: Culicidae) to Spondweni Virus. *PLoS Negl Trop Dis* **10**: e0005083.

Haddow AD, Schuh AJ, Yasuda CY, Kasper MR, Heang V, Huy R, Guzman H, Tesh RB, Weaver SC. 2012. Genetic characterization of Zika virus strains: geographic expansion of the Asian lineage. *PLoS Negl Trop Dis* **6**: e1477.

Haddow AD, Woodall JP. 2016. Distinguishing between Zika and Spondweni viruses. *Bull World Health Organ* **94**: 711–711A.

Haddow AJ, Williams MC, Woodall JP, Simpson DI, Goma LK. 1964. Twelve isolations of zika virus from aedes (stegomyia) africanus (theobald) taken in and above a uganda forest. *Bull World Health Organ* **31**: 57–69.

Hall BG. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol Biol Evol* **22**: 792–802.

Hamer GL, Kitron UD, Brawn JD, Loss SR, Ruiz MO, Goldberg TL, Walker ED. 2008. Culex pipiens (Diptera: Culicidae): a bridge vector of West Nile virus to humans. *J Med Entomol* **45**: 125–128.

Hamer GL, Kitron UD, Goldberg TL, Brawn JD, Loss SR, Ruiz MO, Hayes DB, Walker ED. 2009. Host selection by Culex pipiens mosquitoes and West Nile virus amplification. *Am J Trop Med Hyg* **80**: 268–278.

Hartley WJ, Martin WB, Hakiolu F, Chifney ST. 1969. A viral encephalitis of sheep in Turkey. *Pendik Institute Journal* **2**: 89–100.

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schäffer AA, Brister JR. 2017. Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res* **45**: D482–D490.

Hayasaka D, Ivanov L, Leonova GN, Goto A, Yoshii K, Mizutani T, Kariwa H, Takashima I. 2001. Distribution and characterization of tick-borne encephalitis viruses from Siberia and far-eastern Asia. *J Gen Virol* **82**: 1319–1328.

Henderson BE, Tukei PM, Mccrae AWR, Ssenkubuge Y, Mugo WN. 1970. Virus isolations from ixodid ticks in Uganda. Part II. Kadam virus-a new member of arbovirus group B isolated from Rhipicephalus pravus Donitz. *East African Medical Journal*.

Hirschfeld HO, Wishart J. 1935. A Connection between Correlation and Contingency. *Math Proc Camb Phil Soc* **31**: 520.

Hoshino K, Isawa H, Tsuda Y, Sawabe K, Kobayashi M. 2009. Isolation and characterization of a new insect flavivirus from Aedes albopictus and Aedes flavopictus mosquitoes in Japan. *Virology* **391**: 119–129.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* **24**: 417–441.

Huhtamo E, Cook S, Moureau G, Uzcátegui NY, Sironen T, Kuivanen S, Putkuri N, Kurkela S, Harbach RE, Firth AE, et al. 2014. Novel flaviviruses from mosquitoes: mosquito-specific evolutionary lineages within the phylogenetic group of mosquito-borne flaviviruses. *Virology* **464-465**: 320–329.

Huhtamo E, Moureau G, Cook S, Julkunen O, Putkuri N, Kurkela S, Uzcátegui NY, Harbach RE, Gould EA, Vapalahti O, et al. 2012. Novel insect-specific flavivirus isolated from northern Europe. *Virology* **433**: 471–478.

ICTV. 2017. International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/taxonomy/ (Accessed July 3, 2017).

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* **146**: 1–21.

iTOL. 2017. Interactive Tree Of Life. https://itol.embl.de/ (Accessed August 25, 2017).

Johansen CA, Nisbet DJ, Foley PN, Van Den Hurk AF, Hall RA, Mackenzie JS, Ritchie SA. 2004. Flavivirus isolations from mosquitoes collected from Saibai Island in the Torres Strait, Australia, during an incursion of Japanese encephalitis virus. *Med Vet Entomol* **18**: 281–287.

Johansen CA, Nisbet DJ, Zborowski P, van den Hurk AF, Ritchie SA, Mackenzie JS. 2003. Flavivirus isolations from mosquitoes collected from western Cape York Peninsula, Australia, 1999-2000. *J Am Mosq Control Assoc* **19**: 392–396.

Johansen CA, Williams SH, Melville LF, Nicholson J, Hall RA, Bielefeldt-Ohmann H, Prow NA, Chidlow GR, Wong S, Sinha R, et al. 2017. Characterization of fitzroy river virus and serologic evidence of human and animal infection. *Emerging Infect Dis* **23**: 1289–1299.

Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol* **6**: 494–500.

Karlin S, Blaisdell BE, Schachtel GA. 1990. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. *J Virol* **64**: 4264–4273.

Kay BH, Carley JG, Filippich C. 1975. The multiplication of queensland and new guinean arboviruses in culex annulirostris skuse and aedes vigilax (skuse) (diptera: culicidae). *J Med Entomol* **12**: 279–283.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, et al. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**: 523–526.

Kindhauser MK, Allen T, Frank V, Santhana RS, Dye C. 2016. Zika: the origin and spread of a mosquito-borne virus. *Bull World Health Organ* **94**: 675–686C.

Kofler RM, Hoenninger VM, Thurner C, Mandl CW. 2006. Functional analysis of the tick-borne encephalitis virus cyclization elements indicates major differences between mosquito-borne and tick-borne flaviviruses. *J Virol* **80**: 4099–4113.

Kokernot RH, Smithburn KC, Muspratt J, Hodgson B. 1957. Studies on arthropod-borne viruses of Tongaland. VIII. Spondweni virus, an agent previously unknown, isolated from Taeniorhynchus (Mansonioides) uniformis Theo. *South African Journal of Medical Sciences*.

Kunisawa T, Kanaya S, Kutter E. 1998. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res* **5**: 319–326.

Kuno G. 2007. Host range specificity of flaviviruses: correlation with in vitro replication. *J Med Entomol* **44**: 93–101.

Kuno G, Chang GJ, Tsuchiya KR, Karabatsos N, Cropp CB. 1998. Phylogeny of the genus Flavivirus. *J Virol* **72**: 73–83.

Lanciotti RS, Roehrig JT, Deubel V, Smith J, Parker M, Steele K, Crise B, Volpe KE, Crabtree MB, Scherret JH, et al. 1999. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286**: 2333–2337.

Lara-Ramírez EE, Salazar MI, López-López M de J, Salas-Benito JS, Sánchez-Varela A, Guo X. 2014. Large-scale genomic analysis of codon usage in dengue virus and evaluation of its phylogenetic dependence. *Biomed Res Int* **2014**: 851425.

Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**: 383–397.

Lawrie CH, Uzcátegui NY, Armesto M, Bell-Sakyi L, Gould EA. 2004. Susceptibility of mosquito and tick cell lines to infection with various flaviviruses. *Med Vet Entomol* **18**: 268–274.

Leake CJ, Ussery MA, Nisalak A, Hoke CH, Andre RG, Burke DS. 1986. Virus isolations from mosquitoes collected during the 1982 Japanese encephalitis epidemic in northern Thailand. *Trans R Soc Trop Med Hyg* **80**: 831–837.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**: W242–5.

Li M-H, Fu S-H, Chen W-X, Wang H-Y, Guo Y-H, Liu Q-Y, Li Y-X, Luo H-M, Da W, Duo Ji DZ, et al. 2011. Genotype v Japanese encephalitis virus is emerging. *PLoS Negl Trop Dis* **5**: e1231.

Liehne CG, Leivers S, Stanley NF, Alpers MP, Paul S, Liehne PF, Chan KH. 1976. Ord River arboviruses--isolations from mosquitoes. *Aust J Exp Biol Med Sci* **54**: 499–504.

Liu R, Zhang G, Liu X, Li Y, Zheng Z, Sun X, Yang Y. 2016. Detection of the Siberian Tick-borne Encephalitis Virus in the Xinjiang Uygur Autonomous Region, northwestern China. *Bing Du Xue Bao* **32**: 26–31.

Lloyd S. 1982. Least squares quantization in PCM. *IEEE Trans Inform Theory* **28**: 129–137.

Lobo FP, Mota BEF, Pena SDJ, Azevedo V, Macedo AM, Tauch A, Machado CR, Franco GR. 2009. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One* **4**: e6282.

Lvov DK, Butenko AM, Gromashevsky VL, Kovtunov AI, Prilipov AG, Kinney R, Aristova VA, Dzharkenov AF, Samokhvalov EI, Savage HM, et al. 2004. West Nile virus and other zoonotic viruses in Russia: examples of emerging-reemerging situations. *Arch Virol Suppl* 85–96.

Lvov DK, Neronov VM, Gromashevsky VL, Skvortsova TM, Berezina LK, Sidorova GA, Zhmaeva ZM, Gofman YA, Klimenko SM, Fomina KB. 1976. Karshi" virus, a new flavivirus (Togaviridae) isolated fromOrnithodoros papillipes (Birula, 1895) ticks in Uzbek S.S.R. *Arch Virol* **50**: 29–36.

Ma J-J, Zhao F, Zhang J, Zhou J-H, Ma L-N, Ding Y-Z, Chen H-T, Gu Y-X, Liu Y-S. 2013. Analysis of Synonymous Codon Usage in Dengue Viruses. *J Anim Vet Adv* **12**: 88–98.

Mackenzie JS, Smith DW, Broom AK, Bucens MR. 1993. Australian encephalitis in Western Australia, 1978-1991. *Med J Aust* **158**: 591–595.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1**: 281.

Marshall ID, Woodroofe GM, Hirsch S. 1982. Viruses recovered from mosquitoes and wildlife serum collected in the murray valley of south-eastern australia, february 1974, during an epidemic of encephalitis. *Aust J Exp Biol Med* **60**: 457–470.

Mattingly PF. 1957. Genetical Aspects of the *Aedes aegypti* Problem. *Annals of Tropical Medicine & Parasitology* **51**: 392–408.

May M, Relich RF. 2016. A comprehensive systems biology approach to studying zika virus. *PLoS One* **11**: e0161355.

Meister T, Lussy H, Bakonyi T, Sikutová S, Rudolf I, Vogl W, Winkler H, Frey H, Hubálek Z, Nowotny N, et al. 2008. Serological evidence of continuing high Usutu

virus (Flaviviridae) activity and establishment of herd immunity in wild birds in Austria. *Vet Microbiol* **127**: 237–248.

Melian EB, Hinzman E, Nagasaki T, Firth AE, Wills NM, Nouwens AS, Blitvich BJ, Leung J, Funk A, Atkins JF, et al. 2010. NS1' of flaviviruses in the Japanese encephalitis virus serogroup is a product of ribosomal frameshifting and plays a role in viral neuroinvasiveness. *J Virol* **84**: 1641–1647.

Michely S, Toulza E, Subirana L, John U, Cognat V, Maréchal-Drouard L, Grimsley N, Moreau H, Piganeau G. 2013. Evolution of codon usage in the smallest photosynthetic eukaryotes and their giant viruses. *Genome Biol Evol* **5**: 848–859.

Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. *Viruses* **8**: 66.

Mitchell CJ, Forattini OP, Miller BR. 1986. Vector competence experiments with Rocio virus and three mosquito species from the epidemic zone in Brazil. *Rev Saúde Pública* **20**: 171–177.

Mohammed MAF, Galbraith SE, Radford AD, Dove W, Takasaki T, Kurane I, Solomon T. 2011. Molecular phylogenetic and evolutionary analyses of Muar strain of Japanese encephalitis virus reveal it is the missing fifth genotype. *Infect Genet Evol* **11**: 855–862.

Moratorio G, Iriarte A, Moreno P, Musto H, Cristina J. 2013. A detailed comparative analysis on the overall codon usage patterns in West Nile virus. *Infect Genet Evol* **14**: 396–400.

Mossadegh N, Gissmann L, Müller M, Zentgraf H, Alonso A, Tomakidi P. 2004. Codon optimization of the human papillomavirus 11 (HPV 11) L1 gene leads to increased gene expression and formation of virus-like particles in mammalian epithelial cells. *Virology* **326**: 57–66.

Mugo WN, Shope RE. 1972. Kadam virus: Neutralization studies and laboratory transmission by dermacentor variabilis. *Trans R Soc Trop Med Hyg* **66**: 300–304.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**: 292.

NAMAC. Arboviral diseases and malaria in Australia, 2013–14. *National Arbovirus and Malaria Advisory Committee*. http://www.health.gov.au/internet/main/publishing.nsf/content/cda-arboanrep.htm (Accessed July 20, 2018).

NCBI. 2017. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/ (Accessed January 26, 2018).

NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**: D8–D13.

NCBI Taxonomy browser. 2018. NCBI Taxonomy browser. https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi (Accessed June 18, 2018).

Normile D. 2013. Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts. *Science* **342**: 415.

Nucleotide. 2017. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/nuccore (Accessed June 13, 2017).

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–45.

Omilabu SA, Fagbami AH, Olaleye OD. 1989. Susceptibility of laboratory and domestic animals to experimental infection with Potiskum virus. *Microbios* **60**: 53–58.

Pachler K, Lebl K, Berer D, Rudolf I, Hubalek Z, Nowotny N. 2014. Putative new West Nile virus lineage in Uranotaenia unguiculata mosquitoes, Austria, 2013. *Emerging Infect Dis* **20**: 2119–2122.

PAHO. Zika Cumulative Cases. *Pan American Health Organization*. https://www.paho.org/hq/index.php?option=com_content&view=article&id=12390&Itemid=42090&lang=en (Accessed June 7, 2018).

Pauli G, Bauerfeind U, Blümel J, Burger R, Drosten C, Gröner A, Gürtler L, Heiden M, Hildebrandt M, Jansen B, et al. 2013. West nile virus. *Transfus Med Hemother* **40**: 265–284.

Pearson K. 1901. L III. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6* **2**: 559–572.

Pecorari M, Longo G, Gennari W, Grottola A, Sabbatini AMT, Tagliazucchi S, Savini G, Monaco F, Simone ML, Lelli R, et al. 2009. First human case of Usutu virus neuroinvasive infection, Italy, August-September 2009. *Eurosurveillance*.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, et al. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* **40**: D593–8.

Poidinger M, Hall RA, Mackenzie JS. 1996. Molecular characterization of the Japanese encephalitis serocomplex of the flavivirus genus. *Virology* **218**: 417–421.

Posen HJ, Keystone JS, Gubbay JB, Morris SK. 2016. Epidemiology of Zika virus, 1947–2007. *BMJ Glob Health* **1**: e000087.

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. *Proc Natl Acad Sci USA* **76**: 1697–1701.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.

Puigbò P, Aragonès L, Garcia-Vallvé S. 2010. RCDI/eRCDI: a web-server to estimate codon usage deoptimization. *BMC Res Notes* **3**: 87.

Puigbò P, Bravo IG, Garcia-Vallve S. 2008a. CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* **3**: 38.

Puigbò P, Bravo IG, Garcia-Vallvé S. 2008b. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics* **9**: 65.

Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S. 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res* **35**: W126–31.

Puigbò P, Romeu A, Garcia-Vallvé S. 2008c. HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res* **36**: D524–7.

Ramaiah A, Dai L, Contreras D, Sinha S, Sun R, Arumugaswami V. 2017. Comparative analysis of protein evolution in the genome of pre-epidemic and epidemic Zika virus. *Infect Genet Evol* **51**: 74–85.

Ramakrishna L, Anand KK, Mohankumar KM, Ranga U. 2004. Codon optimization of the tat antigen of human immunodeficiency virus type 1 generates strong immune responses in mice following genetic immunization. *J Virol* **78**: 9174–9189.

RefSeq. 2017. NCBI Reference Sequence Database. https://www.ncbi.nlm.nih.gov/refseq/ (Accessed August 4, 2017).

Rigau-Pérez JG, Clark GG, Gubler DJ, Reiter P, Sanders EJ, Vorndam AV. 1998. Dengue and dengue haemorrhagic fever. *Lancet* **352**: 971–977.

Robin Y, Cornet M, Le Gonidec G, Chateau R, Heme G. 1978. Kedougou virus (Ar D14701): a new arbovirus (flavivirus) isolated in Senegal. *Ann Microbiol (Inst Pasteur)*.

Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**: 2279–2286.

Russell RC. 1998. Mosquito-borne arboviruses in Australia: the current scene and implications of climate change for human health. *Int J Parasitol* **28**: 955–969.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–425.

Sanjuán R, Domingo-Calap P. 2016. Mechanisms of viral mutation. *Cell Mol Life Sci* **73**: 4433–4448.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J Virol* **84**: 9733–9748.

Sayers E. 2008. Entrez Programming Utilities Help. *National Center for Biotechnology In-formation (NCBI)*. https://www.ncbi.nlm.nih.gov/books/NBK25501/ (Accessed April 24, 2018).

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**: D5–15.

Sejvar JJ. 2003. West nile virus: an historical overview. *Ochsner J* **5**: 6–10.

Shackelton LA, Parrish CR, Holmes EC. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J Mol Evol* **62**: 551–563.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**: 28–38.

Sharp PM, Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281–1295.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* **21**: 835–841.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125–5143.

Shope RE. 1963. The use of a microhemagglutination-inhibition test to follow antibody response after arthropod-borne virus infection in a community of forest animals. *An Microbiol* **11**: 167–169.

Simmonds P, Becher P, Bukh J, Gould EA, Meyers G, Monath T, Muerhoff S, Pletnev A, Rico-Hesse R, Smith DB, et al. 2017. ICTV virus taxonomy profile: flaviviridae. *J Gen Virol* **98**: 2–3.

Singapore Zika Study Group. 2017. Outbreak of Zika virus infection in Singapore: an epidemiological, entomological, virological, and clinical analysis. *Lancet Infect Dis* **17**: 813–821.

Singh NK, Tyagi A. 2017. A detailed analysis of codon usage patterns and influencing factors in Zika virus. *Arch Virol* **162**: 1963–1973.

Singh NK, Tyagi A, Kaur R, Verma R, Gupta PK. 2016. Characterization of codon usage pattern and influencing factors in Japanese encephalitis virus. *Virus Res* **221**: 58–65.

Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**: 1122–1125.

Smithburn KC, Haddow AJ. 1951. Ntaya virus; a hitherto unknown agent isolated from mosquitoes collected in Uganda. *Proc Soc Exp Biol Med* **77**: 130–133.

Sokal RR, Michener CD. 1958. A Statistical Method for Evaluating Systematic Relationships. *Kans Univ sci bull* **38**: 1409–1438.

St George TD, Doherty RL, Carley JG, Filippich C, Brescia A, Casals J, Kemp DH, Brothers N. 1985. The isolation of arboviruses including a new flavivirus and a new Bunyavirus from Ixodes (Ceratixodes) uriae (Ixodoidea: Ixodidae) collected at Macquarie Island, Australia, 1975-1979. *Am J Trop Med Hyg* **34**: 406–412.

St George TD, Standfast HA, Doherty RL, Carley JG, Fillipich C, Brandsma J. 1977. The isolation of Saumarez Reef virus, a new flavivirus, from bird ticks Ornithodoros capensis and Ixodes eudyptidis in Australia. *Aust J Exp Biol Med Sci* **55**: 493–499.

Stadler K, Allison SL, Schalich J, Heinz FX. 1997. Proteolytic activation of tick-borne encephalitis virus by furin. *J Virol* **71**: 8475–8481.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Steinhaus H. 1956. Sur la division des corp materiels en parties. *Bull Acad Polon Sci* **1**.

Studier JA, Keppler KJ. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* **5**: 729–731.

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.

Tang Y, Diao Y, Chen H, Ou Q, Liu X, Gao X, Yu C, Wang L. 2015. Isolation and genetic characterization of a tembusu virus strain isolated from mosquitoes in Shandong, China. *Transbound Emerg Dis* **62**: 209–216.

Tatem AJ, Rogers DJ, Hay SI. 2006. Global Transport Networks and Infectious Disease Spread. In *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, Vol. 62 of *Advances in Parasitology*, pp. 293–343, Elsevier.

Taylor LH, Latham SM, Woolhouse ME. 2001. Risk factors for human disease emergence. *Philos Trans R Soc Lond B, Biol Sci* **356**: 983–989.

Traoré-Lamizana M, Fontenille D, Diallo M, Bâ Y, Zeller HG, Mondo M, Adam F, Thonon J, Maïga A. 2001. Arbovirus surveillance from 1990 to 1995 in the Barkedji area (Ferlo) of Senegal, a possible natural focus of Rift Valley fever virus. *J Med Entomol* **38**: 480–492.

Trapido H, Rajagopalan PK, Work TH, Varma MG. 1959. Kyasanur Forest disease. VIII. Isolation of Kyasanur Forest disease virus from naturally infected ticks of the genus Haemaphysalis. *Indian J Med Res* **47**: 133–138.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* **107**: 3645–3650.

Twiddy SS, Holmes EC, Rambaut A. 2003. Inferring the rate and time-scale of dengue virus evolution. *Mol Biol Evol* **20**: 122–129.

van Hemert F, Berkhout B. 2016. Nucleotide composition of the Zika virus RNA genome and its codon usage. *Virol J* **13**: 95.

Wang H, Liang G. 2015. Epidemiology of Japanese encephalitis: past, present, and future prospects. *Ther Clin Risk Manag* **11**: 435–448.

Wang H, Liu S, Zhang B, Wei W. 2016. Analysis of Synonymous Codon Usage Bias of Zika Virus and Its Adaption to the Hosts. *PLoS One* **11**: e0166260.

Varma MG, Webb HE, Pavri KM. 1960. Studies on the transmission of Kyasanur Forest disease virus by Haemaphysalis spinigera Newman. *Trans R Soc Trop Med Hyg* **54**: 509–516.

Vazquez A, Sanchez-Seco MP, Ruiz S, Molero F, Hernandez L, Moreno J, Magallanes A, Tejedor CG, Tenorio A. 2010. Putative new lineage of west nile virus, Spain. *Emerging Infect Dis* **16**: 549–552.

Wengler G, Wengler G, Gross HJ. 1978. Studies on virus-specific nucleic acids synthesized in vertebrate and mosquito cells infected with flaviviruses. *Virology* **89**: 423–437.

WHO. 2018. Dengue and severe dengue. *World Health Organization*. http://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (Accessed June 7, 2018).

Williams RE, Casals J, Moussa MI, Hoogstraal H. 1972. Royal farm virus: a new tick-borne group B agent related to the RSSE complex. *Am J Trop Med Hyg* **21**: 582–586.

Williams, Richard A. J. 2012. Yaoundé-like virus in resident wild bird, Ghana. *Afr J Microbiol Res* **6**.

ViPR. 2017. Virus Pathogen Database and Analysis Resource (ViPR). https://www.viprbrc.org/brc/home.spg?decorator=flavi (Accessed August 22, 2017).

Virus Variation Resource. 2017. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (NCBI). *Virus Variation Resource*. https://www.ncbi.nlm.nih.gov/genome/viruses/variation/ (Accessed August 22, 2017).

Virus-Host DB. 2017. Virus-Host Database. http://www.genome.jp/virushostdb/ (Accessed October 6, 2017).

Wolfe MS, Calisher CH, McGuire K. 1982. Spondweni virus infection in a foreign resident of Upper Volta. *Lancet* **2**: 1306–1308.

Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature* **447**: 279–283.

Wood OL, Moussa MI, Hoogstraal H, Büttiker W. 1982. Kadam Virus (Togaviridae, Flavivirus) Infecting Camel-Parasitizing Hyalomma Dromedarii Ticks (Acari: Ixodidae) in Saudi Arabia12. *J Med Entomol* **19**: 207–208.

Woolhouse M, Scott F, Hudson Z, Howey R, Chase-Topping M. 2012. Human viruses: discovery and emergence. *Philos Trans R Soc Lond B, Biol Sci* **367**: 2864–2871.

Yokoyama S, Starmer WT. 2017. Possible roles of new mutations shared by asian and american zika viruses. *Mol Biol Evol*.

Zaki AM. 1997. Isolation of a flavivirus related to the tick-borne encephalitis complex from human cases in Saudi Arabia. *Trans R Soc Trop Med Hyg* **91**: 179–181.

Zhao K-N, Chen J. 2011. Codon usage roles in human papillomavirus. *Rev Med Virol* **21**: 397–411.

Zhou J, Gao Z, Zhang J, Chen H, Pejsak Z, Ma L, Ding Y, Liu Y. 2012. Comparative [corrected] codon usage between the three main viruses in pestivirus genus and their natural susceptible livestock. *Virus Genes* **44**: 475–481.

Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J Virol* **73**: 4972–4982.

9. APPENDICES

Appendix 1. Summary of data related to optimal host identification

Table 1. List of viruses of the genus *Flavivirus* (13 pages). The table includes the NCBI accession codes of complete coding sequences (CDSs) and amino acid sequences (AAs), guanosine-cytosine composition based on the third nucleotide of each codon (GC3), hosts and classification according to literature, and a summary of results from the correspondence analyses of normalized Codon Adaptation Index (nCAI) values for all 94 flaviviruses used in this thesis (figure 18). These include values of dimension 1 (Dim. 1) and dimension 2 (Dim. 2), host determination based on multivariate normal distribution centroids (confidence level of 0.95) and the nearest host based on Euclidian distances to the centroids of flavivirus subgroups. The hosts (H) and vectors (V) from the literature are conservative, i.e. they are included only by being sequenced from an organism or they were successfully infected in a laboratory setting.

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Aedes flavivirus	NC_012932.1	YP_003029843.1	0.531	*Aedes albopictus* (H) [1] *Aedes flavopictus* (H) (Hoshino et al. 2009)	IOFV	0.395	-1.636	IOFV / Mosquito	*Aedes albopictus*
Alfuy virus	AY898809.1	AAX82481.1	0.525	*Mus musculus* (H) [1] *Centropus phasianinus* (H) [1] *Mammalia* (H) (Doherty et al. 1971) *Culex pullus* (V) (Doherty et al. 1979) *Culex sitiens* (V) (Johansen et al. 2003)	MBFV	-0.025	-0.498	MBFV / Vertebrate, mosquito	*Columba livia*
Alkhurma hemorrhagic fever virus	NC_004355.1	NP_722551.1	0.579	*Ixodes petauristae* (V) (Charrel et al. 2005) *Ixodes ceylonensis* (V) (Charrel et al. 2005) *Homo sapiens* (H) (Charrel et al. 2005; Zaki 1997)	TBFV	0.792	0.745	TBFV / Tick, vertebrate	*Sus scrofa*
Anopheles flavivirus variant 1	NC_031327.1	YP_009305197.1	0.525	*Anopheles gambiae* (H) [1]	IOFV	0.208	-1.993	IOFV / Mosquito	*Aedes albopictus*
Anopheles flavivirus variant 2	KX148547.1	AOR51360.1	0.519	*Anopheles gambiae* (H) [2]	IOFV	-0.037	-1.419	IOFV / Mosquito, vertebrate	*Aedes albopictus*
Apoi virus	NC_003676.1	NP_620045.1	0.501	*Apodemus argenteus* (H) [1]	UVFV	-0.966	0.752	MBFV / Vertebrate	*Homo sapiens*
Bagaza virus	NC_012534.1	YP_002790883.1	0.529	*Culex tritaeniorhynchus* (V) (Bondre et al. 2009) *Homo sapiens* (H) (Bondre et al. 2009) *Alectoris rufa* (H) (Agüero et al. 2011) *Phasianus colchicus* (H) (Agüero et al. 2011)	MBFV	-0.081	0.126	MBFV / Vertebrate, mosquito	*Gallus gallus*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Bainyik virus	KM225264.1	AIJ19433.1	0.528	*Culicidae* (V) [1] *Aedes albopictus* (V) [1] *Mus musculus* (H) [1] *Aedes* sp. (V) [1] Vertebrates (H) (Colmant et al. 2016)	MBFV	-0.093	0.111	MBFV / Vertebrate, mosquito	*Gallus gallus*
Bamaga virus	NC_033725.1	YP_009345036.1	0.500	*Culex sitiens* (V) [1] *Marsupialia* (H) (Colmant et al. 2016)	MBFV	-1.003	0.537	UVFV / Vertebrate	*Alligator mississipiensis*
Banzi virus	DQ859056.1	ABI54472.1	0.548	*Culex rubinotus* (V) (Grard et al. 2010) *Mansonia africana* (V) (Grard et al. 2010) *Mesocricetus auratus* (H) (Grard et al. 2010) *Mastomys natalensis* (H) (Grard et al. 2010)	MBFV	-0.191	1.293	MBFV / Vertebrate	*Myotis brandtii*
Bouboui virus	NC_033693.1	YP_009344961.1	0.493	*Antilocapra* (H) [1] *Rodentia* (H) [1] *Cercopithecus nictitans* (H) [1] *Papio papio* (H) [1] *Anopheles paludis* (V) (Grard et al. 2010) *Eretmapodites inornatus* (V) (Grard et al. 2010) *Aedes* spp. (V) (Grard et al. 2010) *Culex* spp. (V) (Grard et al. 2010)	MBFV	-1.111	0.669	MBFV / Vertebrate	*Alligator mississipiensis*
Bussuquara virus	NC_009026.2	YP_001040004.1	0.521	*Chlorocebus aethiops* (H) [1] *Homo sapiens* (H) [1] *Proechimys* spp. (H) (Shope 1963) *Alouatta belzebul* (H) (Gomes and Causey 1959)	MBFV	-0.200	0.071	MBFV / Vertebrate, mosquito	*Gallus gallus*
Cacipacore virus	NC_026623.1	YP_009126874.1	0.519	*Formicarius analis* (H) [1] *Homo sapiens* (H) (Batista et al. 2011)	MBFV	-0.224	-0.236	MBFV / Vertebrate, mosquito	*Columba livia*
Calbertado virus	KX669689.1	ASA45776.1	0.567	*Culex tarsalis* (H) [2] *Culex pipiens* (H) (Bolling et al. 2011)	IOFV	1.330	-2.695	IOFV / Mosquito	*Aedes albopictus*
Cell fusing agent virus	NC_001564.2	YP_009259257.1	0.566	*Aedes aegypti* (H) [1] *Culicidae* (H) (Cook et al. 2006)	IOFV	0.678	0.253	MBFV / Mosquito, vertebrate, tick	*Bos taurus*
Chaoyang virus	NC_017086.1	YP_005454257.1	0.492	*Culicidae* (H) [1]	dhIOFV	-0.536	-1.048	dhIOFV / Vertebrate, mosquito	*Anas platyrhynchos*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Culex flavivirus	NC_008604.2	YP_899469.2	0.603	*Culex pipiens* (H) [1]	IOFV	1.874	-1.680	IOFV / Mosquito	*Culex quinquefasciatus*
Culiseta flavivirus	NC_030290.1	YP_009256193.1	0.492	*Culiseta melanura* (H) [1]	IOFV	-0.293	-2.071	IOFV / Mosquito	*Aedes aegypti*
Deer tick virus	AF311056.1	AAL32169.1	0.561	*Ixodes scapularis* (H) [1]	TBFV	0.392	0.525	TBFV / Tick, vertebrate	*Bos taurus*
Dengue virus 1	NC_001477.1	NP_059433.1	0.462	*Aedes aegypti* (V) [1] *Aedes albopictus* (V) [1] *Homo sapiens* (H) [1]	MBFV	-1.244	-0.627	dhIOFV / Vertebrate	*Anas platyrhynchos*
Dengue virus 2	NC_001474.2	NP_056776.2	0.459	*Aedes aegypti* (V) [1] *Erythrocebus patas* (H) [1] *Homo sapiens* (H) [1] *Aedes furcifer* (V) [1] *Aedes taylori* (V) [1]	MBFV	-1.475	-0.059	dhIOFV / Vertebrate	*Anas platyrhynchos*
Dengue virus 3	NC_001475.2	YP_001621843.1	0.468	*Erythrocebus patas* (H) [1] *Homo sapiens* (H) [1] *Diceromyia* (V) [1] *Aedimorphus* (V) [1] *Stegomyia* (V) [1]	MBFV	-1.437	0.706	MBFV / Vertebrate	*Alligator mississipiensis*
Dengue virus 4	NC_002640.1	NP_073286.1	0.481	*Aedes aegypti* (V) [1] *Aedes albopictus* (V) [1] *Homo sapiens* (H) [1] *Aedes polynesiensis* (V) [1]	MBFV	-1.176	1.138	MBFV / Vertebrate	*Homo sapiens*
Donggang virus	NC_016997.1	YP_005352889.1	0.484	*Culicidae* (V) [1] *Aedes* sp. (V) [1]	dhIOFV	-1.092	-0.490	dhIOFV / Vertebrate	*Anas platyrhynchos*
Edge Hill virus	NC_030289.1	YP_009256192.1	0.488	*Macropodidae* (H) [1] *Culex annulirostris* (V) (Grard et al. 2010) *Anopheles meraukensis* (V) (Grard et al. 2010) *Aedes vigilax* (V) (Grard et al. 2010)	MBFV	-1.091	0.123	UVFV / Vertebrate	*Alligator mississipiensis*
Entebbe bat virus	NC_008718.1	YP_950477.1	0.567	*Chiroptera* (H) [1]	MBFV	0.426	0.881	MBFV / Vertebrate, tick	*Myotis davidii*
Far Eastern tick-borne encephalitis virus	JX498940.1	AFV41132.1	0.589	*Ixodes persulcatus* (V) (Hayasaka et al. 2001) *Mus musculus* (H) (Hayasaka et al. 2001)	TBFV	0.958	0.646	TBFV / Tick, vertebrate	*Sus scrofa*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Fitzroy River Virus	KM361634.1	AKH03452.1	0.483	*Aedes normanensis* (V) [2] *Anopheles amictus* (V) (Johansen et al. 2017) *Culex annulirostris* (V) (Johansen et al. 2017) *Mammalia* (H) (Johansen et al. 2017) *Aves* (H) (Johansen et al. 2017)	MBFV	-1.367	1.286	MBFV / Vertebrate	*Homo sapiens*
Gadgets Gully virus	NC_033723.1	YP_009345034.1	0.553	*Aves* (H) [1] *Homo sapiens* (H) [1] *Ixodes uriae* (V) (St George et al. 1985)	TBFV	0.515	0.116	TBFV / Tick, vertebrate, mosquito	*Gallus gallus*
Hanko virus	NC_030401.1	YP_009259489.1	0.488	*Culicidae* (H) [1] *Ochlerotatus punctor* (H) (Huhtamo et al. 2012) *Ochlerotatus caspius* (H) (Huhtamo et al. 2012)	IOFV	-0.211	-2.423	IOFV / Mosquito	*Aedes aegypti*
Iguape virus	AY632538.4	AAV34154.1	0.557	Rodents (H) (Coimbra et al. 1993) Sentinel mouse (H) (Coimbra et al. 1993) Marsupials (H) (Coimbra et al. 1993) Birds (H) (Coimbra et al. 1993)	MBFV	0.449	0.341	MBFV / Vertebrate, mosquito, tick	*Bos taurus*
Ilheus virus	NC_009028.2	YP_001040006.1	0.581	*Culex* (V) [1] *Haemagogus* (V) [1] *Psorophora* (V) [1] *Aves* (H) [1] *Homo sapiens* (H) [1] *Sabethes* (V) [1] *Ochlerotatus* (V) [1] *Trichoprosopon* (V) [1]	MBFV	0.643	1.266	MBFV / Vertebrate, tick	*Sus scrofa*
Ilomantsi virus	NC_024805.1	YP_009056847.1	0.476	*Culicidae* (H) [1]	dhIOFV	-0.963	-1.410	dhIOFV / Vertebrate, mosquito	*Xenopus laevis*
Israel turkey meningoencephalomyelitis virus	KC734549.1	AGV15505.1	0.522	*Meleagris gallopavo* (H) [2] *Ochlerotatus caspius* (V) (Braverman et al. 2003) *Culicoides imicola* (V) (Braverman et al. 2003) *Culex pipiens* (V) (Braverman et al. 2003) *Phlebotomus papatasi* (V) (Braverman et al. 2003) *Culicoides distinctipennis* (V) (Braverman et al. 1977)	MBFV	-0.218	-0.028	MBFV / Vertebrate, mosquito	*Gallus gallus*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Japanese encephalitis virus	NC_001437.1	NP_059434.1	0.557	*Culex tritaeniorhynchus* (V) [1] *Ardeidae* (H) [1] *Homo sapiens* (H) [1] *Equus caballus* (H) [1] *Sus scrofa* (H) [1] *Bos Taurus* (H) [1] *Culex gelidus* (V) [1]	MBFV	0.544	0.060	MBFV / Vertebrate, mosquito, tick	*Gallus gallus*
Jugra virus	NC_033699.1	YP_009344969.1	0.491	*Cynopterus brachyotis* (H) [1] *Aedes* sp. (V) (Grard et al. 2010) *Uranotaenia* sp. (V) (Grard et al. 2010)	MBFV	-1.201	0.991	MBFV / Vertebrate	*Homo sapiens*
Jutiapa virus	NC_026620.1	YP_009126871.1	0.447	*Sigmodon hispidus* (H) [1]	UVFV	-1.797	-0.035	UVFV / Vertebrate	*Xenopus laevis*
Kadam virus	NC_033724.1	YP_009345035.1	0.560	*Homo sapiens* (H) [1] *Rhipicephalus pravus* (V) (Henderson et al. 1970) *Rhipicephalus pulchellus* (V) (Davies 1978) *Amblyomma variegatum* (V) (Davies 1978) *Hyalomma dromedariil* (V) (Wood et al. 1982) *Dermacentor variabilis* (V) (Mugo and Shope 1972) *Mus musculus* (H) (Mugo and Shope 1972)	TBFV	0.298	0.923	TBFV / Tick, vertebrate	*Myotis davidii*
Kamiti River virus	NC_005064.1	NP_891560.1	0.541	*Aedes* (H) [1]	IOFV	0.248	-0.732	dhIOFV / Mosquito, vertebrate	*Aedes albopictus*
Karshi virus	NC_006947.1	YP_224133.1	0.608	*Homo sapiens* (H) [1] *Rodentia* (H) [1] *Ornithodoros papillipes* (V) (Lvov et al. 1976) *Mus musculus* (H) (Lvov et al. 1976)	TBFV	1.376	1.387	TBFV / Tick, vertebrate	*Sus scrofa*
Kedougou virus	NC_012533.1	YP_002790882.1	0.595	*Culicidae* (V) [1] *Aedes dalzieli* (V) (Fontenille et al. 1998) *Homo sapiens* (H) (Robin et al. 1978)	MBFV	0.816	1.715	TBFV / Vertebrate	*Sus scrofa*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Kokobera virus	NC_009029.2	YP_001040007.1	0.527	*Aedes albopictus* (V) [1] *Macropus* (H) [1] *Wallabia* (H) [1] *Homo sapiens* (H) [1] *Culex annulirostris* (V) [1] *Ochlerotatus vigilax* (V) [1] *Ochlerotatus camptorhynchus* (V) [1] *Culex sitiens* (V) (Johansen et al. 2004)	MBFV	0.036	0.089	MBFV / Vertebrate, mosquito	*Gallus gallus*
Koutango virus	EU082200.2	ABW76844.2	0.549	*Gerbilliscus kempi* (H) (Fields et al. 2001) *Rhipicephalus* (V) (Fields et al. 2001) *Hyalomma* (V) (Fields et al. 2001) *Ornithodoros* (V) (Fields et al. 2001) *Aedes aegypti* (V) (Coz et al. 1976) *Homo sapiens* (H) (Traoré-Lamizana et al. 2001) *Mastomys* (H) (Traoré-Lamizana et al. 2001) *Lemniscomys striatus* (H) (Traoré-Lamizana et al. 2001)	MBFV	0.196	0.581	MBFV / Vertebrate, tick	*Bos taurus*
Kunjin virus	JX276662.1	AFR66759.1	0.545	*Culex annulirostis* (V) (Marshall et al. 1982) *Aedes tremulus* (V) (Liehne et al. 1976) *Culex australicus* (V) (Russell 1998) *Culex squamosus* (V) (Doherty et al. 1968) *Aedes vigilax* (V) (Kay et al. 1975) *Culex quinquefasciatus* (V) (Russell 1998) *Homo sapiens* (H) (Mackenzie et al. 1993) *Equus* (H) (Frost et al. 2012) Sentinel chicken (H) (NAMAC) *Nycticorax caledonicus* (H) (Boyle et al. 1983) *Culex pseudovishnui* (V) (Bowen et al. 1970) *Anatidae* sp. (H) (Bowen et al. 1970)	MBFV	0.197	0.337	MBFV / Vertebrate, mosquito, tick	*Gallus gallus*
Kyasanur forest disease virus	AY323490.1	AAQ91607.1	0.603	*Homo sapiens* (H) [1] *Semnopithecus entellus* (H) [1] *Haemaphysalis spinigera* (V) (Trapido et al. 1959) *Gallus gallus* (H) (Varma et al. 1960)	TBFV	1.324	0.825	TBFV / Tick, vertebrate	*Ixodes scapularis*
Lammi virus	NC_024806.1	YP_009056848.1	0.518	*Culicidae* (H) [1]	dhIOFV	-0.173	-0.879	dhIOFV / Vertebrate, mosquito	*Columba livia*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Langat virus	NC_003690.1	NP_620108.1	0.592	*Homo sapiens* (H) [1] *Mus* (H) [1] *Ixodes granulatus* (V) (Gordon Smith 1956) *Haemaphysalis papuana* (V) (Bancroft et al. 1976)	TBFV	0.930	1.103	TBFV / Tick, vertebrate	*Sus scrofa*
Louping ill virus	NC_001809.1	NP_044677.1	0.606	*Homo sapiens* (H) [1] *Canis lupus familiaris* (H) [1] *Equus caballus* (H) [1] *Sus scrofa* (H) [1] *Bos taurus* (H) [1] *Ovis aries* (H) [1] *Ixodes ricinus* (V) [1] *Cervinae* (H) [1]	TBFV	1.348	0.758	TBFV / Tick, vertebrate	*Ixodes scapularis*
Meaban virus	NC_033721.1	YP_009345031.1	0.600	*Aves* (H) [1] *Homo sapiens* (H) [1] *Ornithodoros maritimus* (V) (Arnal et al. 2014)	TBFV	1.280	1.020	TBFV / Tick, vertebrate	*Sus scrofa*
Mercadeo virus	NC_027819.1	YP_009164031.1	0.573	*Culex* (H) [1]	IOFV	1.315	-2.140	IOFV / Mosquito	*Aedes albopictus*
Modoc virus	NC_003635.1	NP_619758.1	0.447	*Homo sapiens* (H) [1] *Peromyscus maniculatus* (H) [1]	UVFV	-1.783	-0.187	UVFV / Vertebrate	*Xenopus laevis*
Montana myotis leukoencephalitis virus	NC_004119.1	NP_689391.1	0.415	*Myotis lucifugus* (H) [1]	UVFV	-2.379	-0.290	UVFV / Vertebrate	*Xenopus laevis*
Mosquito flavivirus	NC_021069.1	YP_007877501.1	0.588	*Culex tritaeniorhynchus* (H) [1]	IOFV	1.447	-1.340	IOFV / Mosquito	*Culex quinquefasciatus*
Murray Valley encephalitis virus	NC_000943.1	NP_051124.1	0.493	*Homo sapiens* (H) [1] *Culex annulirostris* (V) [1]	MBFV	-0.637	-0.495	UVFV / Vertebrate, mosquito	*Columba livia*
Naranjal virus	KF917538.1	AIU94742.1	0.516	Sentinel hamster (H) [2]	MBFV	-0.482	0.580	MBFV / Vertebrate	*Homo sapiens*
Negishi virus	KT224355.1	ALP82435.1	0.607	*Homo sapiens* (H) (Ando et al. 1952)	TBFV	1.388	0.664	TBFV / Tick, vertebrate	*Ixodes scapularis*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
New Mapoon virus	NC_032088.1	YP_009328360.1	0.553	*Culicidae* (H) [1] *Culex annulirostris* (H) [1]	MBFV	0.366	0.767	MBFV / Vertebrate, tick	*Bos taurus*
Nounane virus	NC_033715.1	YP_009345019.1	0.531	*Uranotaenia mashonaensis* (H) [2]	dhIOFV	0.048	-1.026	dhIOFV / Vertebrate, mosquito	*Aedes albopictus*
Ntaya virus	NC_018705.3	YP_006846328.2	0.489	*Homo sapiens* (H) [1] *Mus musculus* (H) [1] *Coquillettidia pseudoconopas* (V) (Smithburn and Haddow 1951) *Uranotaenia alboabdominalis* (V) (Smithburn and Haddow 1951) *Culiseta fraseri* (V) (Smithburn and Haddow 1951) *Coquillettidia aurites* (V) (Smithburn and Haddow 1951) *Aedes simpsoni* (V) (Smithburn and Haddow 1951) *Aedes apicoargenteus* (V) (Smithburn and Haddow 1951) *Aedes africanus* (V) (Smithburn and Haddow 1951) *Aedes albomarginatus* (V) (Smithburn and Haddow 1951) *Lutzia tigripes* (V) (Smithburn and Haddow 1951) *Culex poicilipes* (V) (Smithburn and Haddow 1951) *Culex pruina* (V) (Smithburn and Haddow 1951) *Culex moucheti* (V) (Smithburn and Haddow 1951) *Culex* spp. (V) (Smithburn and Haddow 1951)	MBFV	-0.667	-1.116	UVFV / Vertebrate, mosquito	*Anas platyrhynchos*
Ochlerotatus caspius flavivirus	NC_034242.1	YP_009352228.1	0.499	*Ochlerotatus caspius* (H) [1] *Aedes albopictus* (H) (Ferreira et al. 2013)	IOFV	-0.136	-1.812	IOFV / Mosquito	*Aedes aegypti*
Omsk hemorrhagic fever virus	NC_005062.1	NP_878909.1	0.577	*Ixodes* (V) [1] *Homo sapiens* (H) [1] *Ondatra zibethicus* (H) [1] *Dermacentor reticulatus* (V) [1] *Arvicola amphibius* (H) [1]	TBFV	0.716	0.523	TBFV / Tick, vertebrate	*Bos taurus*
Palm Creek virus	NC_033694.1	YP_009344962.1	0.527	*Coquillettidia xanthogaster* (H) [1]	IOFV	0.530	-2.403	IOFV / Mosquito	*Aedes aegypti*
Paraiso Escondido virus	NC_027999.1	YP_009169331.1	0.473	*Psathyromyia abonnenci* (H) [1]	MBFV	-0.896	-2.060	UVFV / Mosquito	*Aedes aegypti*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Phnom Penh bat virus	NC_034007.1	YP_009350101.1	0.444	*Cynopterus brachyotis* (H) [1]	UVFV	-1.762	-0.143	UVFV / Vertebrate	*Xenopus laevis*
Potiskum virus	NC_029054.2	YP_009433741.1	0.477	*Culicidae* (V) [1] *Homo sapiens* (H) [1] *Rodentia* (H) [1] *Gallus gallus domesticus* (H) (Omilabu et al. 1989)	MBFV	-1.425	0.827	MBFV / Vertebrate	*Alligator mississipiensis*
Powassan virus	NC_003687.1	NP_620099.1	0.574	*Ixodes scapularis* (V) [1] *Homo sapiens* (H) [1] *Marmota monax* (H) [1] *Ixodes spinipalpis* (V) [1] *Dermacentor andersoni* (V) [1] *Ixodes cookei* (V) [1] *Lepus americanus* (H) [1]	TBFV	0.693	0.270	TBFV / Tick, vertebrate, mosquito	*Bos taurus*
Quang Binh virus	NC_012671.1	YP_002884239.1	0.586	*Culicidae* (H) [1] *Culex tritaeniorhynchus* (H) [1]	IOFV	1.608	-2.015	IOFV / Mosquito	*Aedes albopictus*
Rio Bravo virus	NC_003675.1	NP_620044.1	0.407	*Homo sapiens* (H) [1] *Eptesicus fuscus* (H) [1] *Tadarida brasiliensis mexicana* (H) [1] *Molossus ater* (H) [1]	UVFV	-2.468	-0.908	UVFV / Vertebrate	*Xenopus laevis*
Rocio virus	AY632542.4	AAV34158.1	0.584	*Homo sapiens* (H) (de Souza Lopes et al. 1978a, 1978b) *Mus musculus* (H) (de Souza Lopes et al. 1978a) *Zonotrichia capensis* (H) (de Souza Lopes et al. 1978a) *Psorophora ferox* (V) (Mitchell et al. 1986; de Souza Lopes et al. 1981) *Aedes scapularis* (V) (Mitchell et al. 1986)	MBFV	0.779	0.963	MBFV / Vertebrate, tick	*Sus scrofa*
Royal Farm virus	DQ235149.1	ABB90673.1	0.585	*Argas hermanni* (V) (Williams et al. 1972) *Cricetidae* (H) (Williams et al. 1972)	TBFV	1.186	-0.021	TBFV / Tick, vertebrate, mosquito	*Anopheles gambiae*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Saboya virus	NC_033697.1	YP_009344967.1	0.477	*Mus musculus* (H) [1] *Jaculus jaculus* (H) [1] *Arvicanthis niloticus* (H) [1] *Mastomys sp.* (H) [1] *Gerbilliscus kempi* (H) [1] *Phlebotomus duboscqi* (V) (Ba et al. 1999; Fontenille et al. 1994) *Sergentomyia inermis* (V) (Fontenille et al. 1994) *Sergentomyia squamipleuris* (V) (Fontenille et al. 1994) *Sergentomyia adleri* (V) (Fontenille et al. 1994) *Sergentomyia clydei* (V) (Fontenille et al. 1994) *Sergentomyia antennata* (V) (Fontenille et al. 1994) *Sergentomyia buxtoni* (V) (Fontenille et al. 1994) *Sergentomyia dubia* (V) (Fontenille et al. 1994) *Sergentomyia schwetzi* (V) (Fontenille et al. 1994) *Sergentomyia magna* (V) (Fontenille et al. 1994)	MBFV	-1.389	0.636	MBFV / Vertebrate	*Alligator mississipiensis*
Saumarez Reef virus	NC_033726.1	YP_009345037.1	0.576	*Aves* (H) [1] *Homo sapiens* (H) [1] *Ornithodoros capensis* (V) (St George et al. 1977) *Ixodes eudyptidis* (V) (St George et al. 1977)	TBFV	1.063	-0.027	TBFV / Tick, vertebrate, mosquito	*Anopheles gambiae*
Sepik virus	NC_008719.1	YP_950478.1	0.483	*Culicidae* (V) [1] *Ovis aries* (H) (Grard et al. 2010) *Homo sapiens* (H) (Grard et al. 2010)	MBFV	-1.280	0.765	MBFV / Vertebrate	*Alligator mississipiensis*
Siberian tick-borne encephalitis virus	L40361.3	AAF82240.2	0.607	*Ixodes persulcatus* (V) (Liu et al. 2016)	TBFV	1.414	0.244	TBFV / Tick, vertebrate	*Ixodes scapularis*
Sokoluk virus	NC_026624.1	YP_009126875.1	0.588	*Pipistrellus pipistrellus* (H) [1]	MBFV	0.932	0.597	TBFV / Vertebrate, tick	*Ixodes scapularis*
Spanish goat encephalitis virus	NC_027709.1	YP_009162613.1	0.610	*Capra hircus* (H) [1]	TBFV	1.449	0.635	TBFV / Tick, vertebrate	*Sus scrofa*
Spondweni virus	NC_029055.1	YP_009222008.1	0.580	*Aedes circumluteolus* (V) [1] *Mansonia uniformis* (V) (Kokernot et al. 1957) *Homo sapiens* (H) (Wolfe et al. 1982; Haddow et al. 1964)	MBFV	0.794	1.072	MBFV / Vertebrate, tick	*Sus scrofa*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
St. Louis encephalitis virus	NC_007580.2	YP_001008348.1	0.524	*Culex quinquefasciatus* (V) [1] *Dromaius novaehollandiae* (H) [1] *Dasypodidae* (H) [1] *Homo sapiens* (H) [1] *Culex nigripalpus* (V) [1] *Passer domesticus* (H) [1]	MBFV	-0.052	-0.494	MBFV / Vertebrate, mosquito	*Columba livia*
Stratford virus	KM225263.1	AIJ19432.1	0.529	*Aedes albopictus* (V) [1] *Macropodidae* (H) [1] *Homo sapiens* (H) [1] *Equus caballus* (H) [1]	MBFV	0.048	-0.146	TBFV / Vertebrate, mosquito	*Gallus gallus*
Tembusu virus	NC_015843.2	YP_004734464.1	0.509	*Anser* sp. (H) [1] *Culex tritaeniorhynchus* (V) (Leake et al. 1986) *Culex vishnui* (V) (Leake et al. 1986) *Culex gelidus* (V) (Leake et al. 1986) *Culex pipiens* (V) (Tang et al. 2015)	MBFV	-0.518	0.030	UVFV / Vertebrate, mosquito	*Columba livia*
T'Ho virus	NC_034151.1	YP_009351820.1	0.527	*Culex quinquefasciatus* (V) [1] Vertebrates (H) (Farfan-Ale et al. 2009)	MBFV	-0.356	0.559	MBFV / Vertebrate	*Homo sapiens*
Torres virus	KM225265.1	AIJ19434.1	0.526	*Culicidae* (V) [1] *Aedes albopictus* (V) [1] *Culex gelidus* (V) (Johansen et al. 2004) *Sus* (H) (Johansen et al. 2004)	MBFV	-0.141	0.306	TBFV / Vertebrate	*Gallus gallus*
Turkish sheep encephalitis virus	DQ235151.1	ABB90675.1	0.613	*Ovis* (H) (Hartley et al. 1969)	TBFV	1.465	0.582	TBFV / Tick, vertebrate	*Ixodes scapularis*
Tyuleniy virus	NC_023424.1	YP_009001464.1	0.579	*Ixodes uriae* (V) [1]	TBFV	1.058	-0.047	TBFV / Tick, vertebrate, mosquito	*Anopheles gambiae*
Uganda S virus	NC_033698.1	YP_009344968.1	0.464	*Mus musculus* (H) [1] *Saxicola rubetra* (H) [1] *Aedes longipalpis* (V) (Dick and Haddow 1952) *Aedes ingrami* (V) (Dick and Haddow 1952) *Aedes natronius* (V) (Dick and Haddow 1952) *Macaca mulatta* (H) (Dick and Haddow 1952)	MBFV	-1.442	-0.122	UVFV / Vertebrate	*Anas platyrhynchos*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classification	Dim. 1	Dim. 2	Centroid classification	Nearest host
Usutu virus	NC_006551.1	YP_164264.1	0.551	*Aedes albopictus* (V) [1] *Culex pipiens* (V) [1] *Turdus merula* (H) [1] *Homo sapiens* (H) [1] *Anopheles maculipennis* (V) [1] *Ochlerotatus caspius* (V) [1] *Coquillettidia aurites* (V) [1] *Mansonia Africana* (V) [1] *Culex neavei* (V) [1] *Culex perexiguus* (V) [1]	MBFV	0.325	0.497	MBFV / Vertebrate, tick	*Bos taurus*
Wesselsbron virus	NC_012735.1	YP_002922020.1	0.477	*Aedes* (V) [1] *Homo sapiens* (H) [1] *Capra hircus* (H) [1] *Ovis aries* (H) [1]	MBFV	-1.324	0.237	MBFV / Vertebrate	*Alligator mississipiensis*
West Nile virus lineage 1	NC_009942.1	YP_001527877.1	0.560	*Aedes* (V) [1] *Aves* (H) [1] *Homo sapiens* (H) [1] *Amblyomma variegatum* (V) [1] *Hyalomma marginatum* (V) [1] *Rhipicephalus* (V) [1] *Culex* (V) [1] *Mansonia uniformis* (V) [1] *Mimomyia* (V) [1] *Chlorocebus aethiops* (H) [1] *Mesocricetus auratus* (H) [1] *Bubo scandiacus* (H) [1] *Mus* (H) [1] *Corvidae* (H) (Lanciotti et al. 1999)	MBFV	0.463	0.248	MBFV / Vertebrate, mosquito, tick	*Bos taurus*
West Nile virus lineage 2	NC_001563.2	NP_041724.2	0.551	*Homo sapiens* (H) [1]	MBFV	0.341	0.112	MBFV / Vertebrate, mosquito, tick	*Gallus gallus*
Western tick-borne encephalitis virus	NC_001672.1	NP_043135.1	0.595	*Homo sapiens* (H) [1] *Mus musculus* (H) [1] *Ixodes ricinus* (V) [1] *Ixodes persulcatus* (V) [1]	TBFV	1.245	0.276	TBFV / Tick, vertebrate	*Ixodes scapularis*

Virus name	Accession codes		GC3	Based on the literature		Correspondence Analysis (CA)			
	CDS	AA		Putative hosts (H) and vectors (V)	Classifi-cation	Dim. 1	Dim. 2	Centroid classifica-tion	Nearest host
Yaounde virus	NC_034018.1	YP_009350103.1	0.545	*Culex nebulosus*(V) [1] *Culex telesilla* (V) (CDC 1985) *Culex quiarti* (V) (CDC 1985) *Eretmapodites oedipodeios* (V) (CDC 1985) *Aedes aegypti* (V) (CDC 1985) *Culex perfuscus* (V) (CDC 1985) *Culex pruina* (V) (CDC 1985) *Culex duttoni* (V) (CDC 1985) *Bycanistes sharpie* (H) (CDC 1985) *Aves* (H) (Williams, Richard A. J 2012) *Praomys* (H) (CDC 1985)	MBFV	0.338	-0.622	dhIOFV / Vertebrate, mosquito	*Gallus gallus*
Yellow fever virus	NC_002031.1	NP_041726.1	0.537	*Aedes aegypti* (V) [1] *Aedes simpsoni* (V) [1] *Homo sapiens* (H) [1] *Aedes luteocephalus* (V) [1] *Simiiformes* (H) [1]	MBFV	-0.409	1.879	MBFV / Ver-tebrate	*Mus musculus*
Yokose virus	NC_005039.1	NP_872627.1	0.485	*Miniopterus fuliginosus* (H) [1] *Culicidae* (V) (Tajima et al. 2005)	MBFV	-1.119	0.002	MBFV / Ver-tebrate	*Alligator mississip-piensis*
Zika virus	NC_012532.1	YP_002790881.1	0.541	*Aedes aegypti* (V) [1] *Aedes albopictus* (V) [1] *Macaca mulatta* (H) [1] *Homo sapiens* (H) [1] *Mus musculus* (H) [1]	MBFV	-0.131	1.189	MBFV / Ver-tebrate	*Myotis brandtii*

[1] Information obtained from Virus-Host Database (Mihara et al. 2016).
[2] Information obtained from GenBank (Benson et al. 2013).
MBFV = Mosquito-borne flavivirus
TBFV = Tick-borne flavivirus
IOFV = Insect-only flavivirus
UVFV = Unknown vector flavivirus
dhIOFV = Dual-host IOFV
GC3: Proportion of Guanine+Cytosine at the third position of the codon (%)

Appendix 2. Supplementary material related to Zika virus pilot study

Table 1. Codon reference tables for optimal host identification (4 pages). Tables were either self-calculated with Perl scripts or obtained from Codon Usage Database (marked with an asterisk) (Codon Usage Database 2017; Nakamura et al. 2000). The tables were calculated for animals that had at least 10,000 complete coding sequences to ensure representativeness. There were two animals that had less sequences, wild boar (*Sus scrofa*) with 2,953 CDSs and the red junglefowl (*Gallus gallus*) with 6,017 CDSs.

	UUU	UCU	UAU	UGU	UUC	UCC	UAC	UGC	UUA	UCA	UAA	UGA	UUG	UCG	UAG	UGG
Aedes aegypti	105,803	67,316	90,691	70,139	205,867	122,844	160,830	89,653	48,482	75,347	8,409	7,820	162,093	148,951	5,239	82,534
Aedes albopictus	97,145	56,914	79,108	63,514	200,599	123,298	163,078	87,769	38,157	65,567	6,720	5,873	153,673	154,171	4,553	80,333
Alligator mississippiensis	285,736	266,501	210,156	170,626	299,463	254,351	242,753	208,569	143,195	222,283	9,468	15,055	237,164	70,524	7,528	204,449
Anas platyrhynchos	153,243	135,585	95,599	88,295	149,230	126,705	126,168	110,527	73,660	112,384	1,521	2,599	115,508	37,472	1,122	100,505
Anopheles gambiae	94,614	32,112	52,725	50,824	165,698	108,606	173,108	91,618	26,775	43,545	7,161	6,391	75,745	182,093	4,436	75,670
Bos taurus	176,217	159,245	118,803	110,703	230,161	202,850	170,291	140,579	79,353	126,638	8,272	17,113	136,301	59,386	6,476	137,993
Columba livia	112,853	92,944	74,363	65,581	113,012	93,945	97,512	76,376	50,686	75,022	3,715	4,794	84,780	28,616	2,286	73,187
Culex quinquefasciatus	123,786	42,919	50,760	60,923	198,511	134,088	195,233	107,214	25,108	52,078	7,306	6,787	134,792	188,984	4,659	86,930
Gallus gallus*	45,768	38,296	32,211	23,851	54,936	42,683	48,342	36,075	19,129	31,442	2,046	2,986	34,146	14,079	1,281	32,616
Homo sapiens	189,379	171,196	132,715	117,458	216,388	196,012	161,579	135,489	86,682	139,095	5,404	9,518	143,507	50,249	4,328	134,648
Ixodes scapularis	74,374	55,722	31,434	33,921	155,979	105,861	129,415	100,447	19,307	37,110	4,241	8,274	66,146	96,803	4,231	72,965
Mus musculus*	422,153	398,250	298,518	279,729	535,439	444,041	394,074	301,384	165,150	289,799	23,403	40,148	329,668	103,815	19,126	306,619
Myotis brandtii	149,256	134,443	103,108	91,596	189,819	173,931	143,371	113,452	69,465	105,333	5,078	9,823	118,397	45,721	4,583	113,559
Myotis davidii	113,232	103,442	76,640	70,839	155,462	144,783	117,529	95,128	51,237	81,258	3,784	8,156	90,795	39,444	3,674	92,755
Sus scrofa*	18,160	14,246	12,717	10,902	27,973	21,586	22,023	16,776	6,442	10,448	883	1,739	13,518	5,607	628	17,440
Xenopus laevis	475,411	435,671	348,838	278,073	343,215	321,198	290,496	239,628	259,966	343,805	19,075	17,534	332,659	77,483	9,401	245,817

	CUU	CCU	CAU	CGU	CUC	CCC	CAC	CGC	CUA	CCA	CAA	CGA	CUG	CCG	CAG	CGG
Aedes aegypti	81,073	70,190	92,779	70,386	92,845	78,473	111,503	74,523	69,643	116,049	154,187	94,417	248,006	143,421	198,592	92,034
Aedes albopictus	73,743	63,435	85,963	66,043	91,059	79,507	115,067	74,949	67,101	105,136	142,086	89,873	262,230	158,240	205,119	103,979
Alligator mississippiensis	234,983	286,349	199,397	80,956	261,744	253,141	224,817	142,817	134,363	313,100	247,331	88,040	592,915	94,622	565,858	148,393
Anas platyrhynchos	122,813	143,338	94,079	42,336	135,241	121,613	119,654	60,556	60,117	145,719	122,456	45,366	278,261	46,915	271,533	61,272
Anopheles gambiae	51,540	33,672	71,298	60,653	110,524	79,021	127,400	144,786	51,475	75,633	91,709	56,430	341,455	213,374	255,693	126,863
Bos taurus	137,259	183,974	106,989	47,991	226,797	238,009	174,809	122,283	69,555	174,268	124,199	67,416	452,888	98,259	381,244	137,128
Columba livia	83,984	95,097	66,456	29,591	101,401	92,973	87,675	45,990	38,718	99,227	82,943	32,668	206,622	37,795	189,692	48,921
Culex quinquefasciatus	64,004	42,642	58,083	58,474	126,302	97,230	150,387	115,735	44,253	87,713	119,463	75,642	342,737	212,800	255,462	143,812
Gallus gallus*	33,708	41,672	25,885	14,682	45,753	46,097	39,081	28,305	16,211	42,767	33,018	14,339	104,699	21,091	88,743	26,453
Homo sapiens	147,569	198,345	123,609	49,921	212,802	225,420	168,062	115,976	79,488	192,119	140,427	68,859	437,308	81,354	387,120	129,331
Ixodes scapularis	66,042	55,208	36,505	40,240	159,231	120,308	116,979	101,295	31,749	59,161	59,065	44,503	233,226	106,684	169,406	86,673
Mus musculus*	329,757	450,637	260,637	114,854	495,018	446,868	375,626	229,758	198,032	423,707	293,318	161,412	969,515	151,521	836,320	250,836
Myotis brandtii	114,062	157,094	94,505	38,894	185,980	195,453	149,357	88,661	60,061	148,475	111,991	56,543	375,305	64,984	332,354	107,579
Myotis davidii	86,828	124,280	72,079	30,930	154,989	167,094	125,963	77,771	45,385	116,015	84,072	44,250	317,889	60,214	271,506	94,037
Sus scrofa*	13,109	18,561	9,900	4,792	27,053	25,796	18,267	14,145	6,653	16,692	11,567	6,519	53,901	9,860	40,912	13,943
Xenopus laevis	378,042	377,145	298,817	119,961	274,223	244,929	267,467	120,706	213,125	432,029	395,345	122,409	540,897	91,856	585,594	119,667

	AUU	ACU	AAU	AGU	AUC	ACC	AAC	AGC	AUA	ACA	AAA	AGA	AUG	ACG	AAG	AGG
Aedes aegypti	152,166	86,616	166,027	105,612	207,964	148,111	221,033	123,334	76,072	84,025	220,087	53,146	189,682	133,677	268,281	43,987
Aedes albopictus	137,495	76,150	147,062	100,974	208,043	153,503	225,036	125,677	67,644	75,402	206,848	47,629	186,399	137,826	273,900	44,683
Alligator mississippiensis	289,540	245,848	311,827	221,332	318,302	263,582	322,932	326,064	160,920	294,026	477,989	230,179	377,994	97,321	532,404	200,385
Anas platyrhynchos	141,066	117,055	148,135	106,071	153,436	121,061	168,842	167,445	80,265	148,079	247,562	118,152	173,849	57,102	251,100	107,972
Anopheles gambiae	92,725	39,481	100,091	64,028	202,044	140,594	219,692	156,991	61,791	59,333	126,145	23,006	166,014	201,293	262,509	21,939
Bos taurus	159,826	133,607	163,964	129,267	234,653	210,799	212,732	224,422	76,928	150,192	251,908	130,360	233,655	79,271	350,736	134,871
Columba livia	104,393	80,840	107,072	73,499	121,220	92,966	128,058	113,525	58,143	104,091	182,096	84,864	131,436	42,191	189,730	74,351
Culex quinquefasciatus	128,311	53,265	99,723	87,381	246,469	173,860	278,300	147,234	39,912	52,962	149,090	36,161	187,954	191,512	336,380	40,696
Gallus gallus*	45,653	36,078	46,039	30,390	59,906	44,951	61,099	54,867	23,805	43,884	74,256	33,289	62,972	20,943	93,393	31,945
Homo sapiens	175,259	147,134	190,114	139,465	220,634	203,602	206,688	220,505	82,922	167,136	278,169	133,268	236,510	66,200	353,825	131,616
Ixodes scapularis	57,705	45,993	45,690	42,553	134,295	111,797	147,940	121,031	31,509	54,105	78,016	39,079	127,674	107,148	208,833	78,311
Mus musculus*	377,698	335,039	382,284	311,331	552,184	465,115	499,149	483,013	180,467	391,437	537,723	297,135	559,953	138,180	825,270	299,472
Myotis brandtii	139,814	114,452	148,430	113,992	199,193	180,990	185,755	185,631	66,531	131,463	226,768	109,231	204,831	65,277	313,098	118,086
Myotis davidii	105,319	86,591	110,660	87,677	163,717	149,818	150,586	156,457	49,298	100,435	169,437	82,710	162,715	57,159	249,402	96,344
Sus scrofa*	15,660	13,038	16,598	11,082	28,754	26,456	25,606	23,327	7,191	14,403	23,681	12,052	25,610	9,022	38,617	13,258
Xenopus laevis	470,669	372,157	520,334	336,460	352,434	296,177	419,539	337,006	300,000	463,817	720,130	342,826	502,530	95,128	601,326	252,104

	GUU	GCU	GAU	GGU	GUC	GCC	GAC	GGC	GUA	GCA	GAA	GGA	GUG	GCG	GAG	GGG
Aedes aegypti	128,092	123,462	238,605	116,851	115,996	177,203	192,964	115,501	75,096	112,842	295,762	166,203	173,024	103,277	220,013	58,749
Aedes albopictus	120,907	114,500	230,285	116,327	117,235	184,026	202,864	117,142	71,956	106,383	283,652	156,762	180,844	114,481	230,877	62,953
Alligator mississippiensis	220,062	336,797	425,108	182,259	214,803	357,261	392,225	290,968	146,333	328,225	560,811	297,308	410,333	94,451	643,968	250,922
Anas platyrhynchos	117,470	181,596	201,191	94,731	105,498	153,253	180,425	133,781	74,738	173,629	275,108	151,883	198,074	45,370	284,768	115,330
Anopheles gambiae	64,418	75,543	173,980	125,515	109,118	190,816	212,737	204,626	60,279	106,001	179,852	87,574	229,677	195,713	283,970	78,940
Bos taurus	113,850	193,365	220,484	111,861	168,797	326,809	288,879	256,275	70,781	160,241	302,609	174,863	312,161	100,065	441,648	185,928
Columba livia	84,579	119,203	148,911	69,163	84,576	113,404	141,019	95,801	50,144	110,554	201,781	109,153	151,071	39,420	213,638	86,177
Culex quinquefasciatus	113,170	86,702	170,362	108,252	153,568	225,445	270,497	167,341	47,010	81,107	221,992	143,559	218,927	184,999	313,182	89,461
Gallus gallus*	35,593	56,528	68,683	30,898	36,917	62,202	67,783	53,631	21,277	51,713	84,178	47,765	76,624	24,768	111,123	43,513
Homo sapiens	121,302	204,091	246,943	117,456	155,761	311,996	278,549	247,607	78,882	178,106	336,665	183,190	305,878	84,501	451,726	182,999
Ixodes scapularis	56,833	73,718	66,403	60,167	135,465	195,993	232,841	161,342	31,338	81,570	121,651	88,146	190,498	115,433	225,468	82,020
Mus musculus*	262,535	491,093	515,049	280,522	377,902	637,878	638,504	520,069	182,733	388,723	661,498	411,344	696,158	157,124	965,963	372,099
Myotis brandtii	97,606	165,822	196,915	92,699	140,631	262,234	251,134	199,899	60,268	140,268	275,392	144,558	268,064	65,575	386,908	155,884
Myotis davidii	74,166	130,534	149,620	72,558	116,940	225,972	209,517	171,189	45,056	110,134	205,844	112,003	222,467	60,004	317,761	132,602
Sus scrofa*	10,755	19,642	22,309	11,651	20,245	36,983	33,278	29,987	6,447	15,141	27,310	18,686	38,572	10,339	48,046	21,555
Xenopus laevis	351,271	411,579	613,288	253,098	244,083	319,491	432,347	265,240	240,447	424,584	772,321	441,213	434,149	85,444	662,177	270,557

Table 2. Accession codes for complete coding sequences of Zika viruses used for pilot study. Viruses were grouped based on genotype and annotated based on the geographic location they were found. Spondweni virus (in bolded font) was used as an outgroup.

Code	Genotype	Continent/Annotation	Code	Genotype	Continent/Annotation
KU820898	Asian	Asia1	KX156774	Asian	NorthAmerica42
KU740184	Asian	Asia2	KX156775	Asian	NorthAmerica43
KU820899	Asian	Asia3	KX156776	Asian	NorthAmerica44
KU761564	Asian	Asia4	KX198135	Asian	NorthAmerica45
KU681082	Asian	Asia5	KX446950	Asian	NorthAmerica46
KU681081	Asian	Asia6	KX446951	Asian	NorthAmerica47
KU744693	Asian	Asia7	KX694534	Asian	NorthAmerica48
KU955589	Asian	Asia8	KY120348	Asian	NorthAmerica49
KU955590	Asian	Asia9	KX369547	Asian	Oceania1
KU955593	Asian	Asia10	KX447509	Asian	Oceania2
KU963796	Asian	Asia11	KX447510	Asian	Oceania3
KX056898	Asian	Asia12	KX447511	Asian	Oceania4
KX117076	Asian	Asia13	KX447512	Asian	Oceania5
KX185891	Asian	Asia14	KX447513	Asian	Oceania6
KX253996	Asian	Asia15	KX447514	Asian	Oceania7
KU866423	Asian	Asia16	KX447515	Asian	Oceania8
KX377336	Asian	Asia17	KX447516	Asian	Oceania9
KX601167	Asian	Asia18	KX447517	Asian	Oceania10
KX813683	Asian	Asia19	KJ776791	Asian	Oceania11
KX827309	Asian	Asia20	KX806557	Asian	Oceania12
LC190723	Asian	Asia21	KU321639	Asian	SouthAmerica1
LC191864	Asian	Asia22	KU729217	Asian	SouthAmerica2
KX694532	Asian	Asia23	KU729218	Asian	SouthAmerica3
KX694533	Asian	Asia24	KU497555	Asian	SouthAmerica4
KU853013	Asian	Europe1	KU707826	Asian	SouthAmerica5
KU991811	Asian	Europe2	KU527068	Asian	SouthAmerica6
KX673530	Asian	Europe3	KU365780	Asian	SouthAmerica7
KY003153	Asian	Europe4	KU365779	Asian	SouthAmerica8
KY003154	Asian	Europe5	KU365778	Asian	SouthAmerica9
KU922960	Asian	NorthAmerica1	KU365777	Asian	SouthAmerica10
KU922923	Asian	NorthAmerica2	KU312312	Asian	SouthAmerica11
KU647676	Asian	NorthAmerica3	KU940228	Asian	SouthAmerica12
KU501217	Asian	NorthAmerica4	KU937936	Asian	SouthAmerica13
KU501216	Asian	NorthAmerica5	KX197192	Asian	SouthAmerica14
KU501215	Asian	NorthAmerica6	KX247646	Asian	SouthAmerica15
KU870645	Asian	NorthAmerica7	KX280026	Asian	SouthAmerica16
KU509998	Asian	NorthAmerica8	KU758877	Asian	SouthAmerica17
KX051563	Asian	NorthAmerica9	KX520666	Asian	SouthAmerica18
KX247632	Asian	NorthAmerica10	KX548902	Asian	SouthAmerica19
KX262887	Asian	NorthAmerica11	KX702400	Asian	SouthAmerica20
KX377337	Asian	NorthAmerica12	KU820897	Asian	SouthAmerica21
KX601168	Asian	NorthAmerica13	KX197205	Asian	SouthAmerica22
KX766029	Asian	NorthAmerica14	KX811222	Asian	SouthAmerica23
KX832731	Asian	NorthAmerica15	KX879603	Asian	SouthAmerica24
KX856011	Asian	NorthAmerica16	KX879604	Asian	SouthAmerica25
KX838905	Asian	NorthAmerica17	KY014296	Asian	SouthAmerica26
KX838906	Asian	NorthAmerica18	KR872956	Asian	SouthAmerica27
KX842449	Asian	NorthAmerica19	KX087102	Asian	SouthAmerica28
KX922703	Asian	NorthAmerica20	HQ234498	East African	EastAfrica1
KX922704	Asian	NorthAmerica21	KU720415	East African	EastAfrica2
KX922707	Asian	NorthAmerica22	KF268949	East African	EastAfrica3
KX922708	Asian	NorthAmerica23	KF268948	East African	EastAfrica4
KY014295	Asian	NorthAmerica24	LC002520	East African	EastAfrica5
KY014299	Asian	NorthAmerica25	KF383119	East African	EastAfrica6
KY014300	Asian	NorthAmerica26	AY632535	East African	EastAfrica7
KY014303	Asian	NorthAmerica27	DQ859059	East African	EastAfrica8
KY014304	Asian	NorthAmerica28	KU955594	East African	EastAfrica9
KY014314	Asian	NorthAmerica29	KX377335	East African	EastAfrica10
KY014321	Asian	NorthAmerica30	KX601169	East African	EastAfrica11
KY014322	Asian	NorthAmerica31	KX830960	East African	EastAfrica12
KY014323	Asian	NorthAmerica32	KU963573	East African	EastAfrica13
KY014324	Asian	NorthAmerica33	HQ234500	West African	WestAfrica1
KY075932	Asian	NorthAmerica34	KU955591	West African	WestAfrica2
KY075933	Asian	NorthAmerica35	KU955592	West African	WestAfrica3
KY075934	Asian	NorthAmerica36	KU955595	West African	WestAfrica4
KY075935	Asian	NorthAmerica37	KX601166	West African	WestAfrica5
KY075936	Asian	NorthAmerica38	KU963574	West African	WestAfrica6
KY075937	Asian	NorthAmerica39	KX198134	West African	WestAfrica7
KY075938	Asian	NorthAmerica40	**DQ859064**	**–**	**Spondwenivirus**
KX087101	Asian	NorthAmerica41			

Table 3. Average and standard deviation (SD) values of relative synonymous codon usage (RSCU) values in Zika viruses from different continents (2 pages). The codon usages are similar across different continents but show minor differences in usage biases. The amino acids (AAs) encoded by synonymous codons are marked in one-letter abbreviations.

		East Africa		West Africa		Asia		America	
Codon	**AA**	**Average**	**SD**	**Average**	**SD**	**Average**	**SD**	**Average**	**SD**
TTT	**F**	0.996	0.032	0.940	0.008	0.998	0.031	1.011	0.017
TTC	**F**	1.004	0.032	1.060	0.008	1.002	0.031	0.989	0.017
TTA	**L**	0.390	0.021	0.275	0.009	0.312	0.019	0.309	0.010
TTG	**L**	1.319	0.023	1.391	0.029	1.289	0.038	1.325	0.019
CTT	**L**	0.867	0.029	0.844	0.036	0.800	0.058	0.754	0.025
CTC	**L**	1.030	0.010	0.979	0.030	0.977	0.037	1.007	0.020
CTA	**L**	0.595	0.025	0.721	0.010	0.678	0.018	0.688	0.024
CTG	**L**	1.799	0.028	1.790	0.043	1.945	0.042	1.917	0.022
ATT	**I**	0.811	0.027	0.733	0.027	0.888	0.055	0.897	0.016
ATC	**I**	1.286	0.048	1.310	0.028	1.126	0.039	1.120	0.019
ATA	**I**	0.903	0.027	0.957	0.005	0.986	0.020	0.983	0.012
GTT	**V**	0.817	0.023	0.821	0.009	0.853	0.035	0.852	0.012
GTC	**V**	1.025	0.009	0.974	0.013	1.133	0.054	1.135	0.016
GTA	**V**	0.433	0.049	0.411	0.013	0.371	0.021	0.376	0.011
GTG	**V**	1.726	0.040	1.794	0.009	1.643	0.035	1.637	0.010
TCT	**S**	0.873	0.055	0.660	0.011	0.856	0.028	0.886	0.017
TCC	**S**	0.898	0.066	1.126	0.005	0.995	0.025	0.974	0.015
TCA	**S**	1.486	0.019	1.401	0.008	1.530	0.020	1.544	0.009
TCG	**S**	0.526	0.028	0.584	0.017	0.385	0.019	0.368	0.010
AGT	**S**	1.079	0.055	1.100	0.016	0.961	0.021	0.944	0.021
AGC	**S**	1.139	0.053	1.130	0.013	1.273	0.022	1.284	0.021
CCT	**P**	0.765	0.052	0.665	0.028	0.669	0.041	0.651	0.020
CCC	**P**	1.060	0.048	1.155	0.014	1.131	0.045	1.136	0.019
CCA	**P**	1.860	0.013	1.864	0.014	1.756	0.054	1.785	0.013
CCG	**P**	0.315	0.017	0.316	0.055	0.445	0.038	0.428	0.013
ACT	**T**	0.932	0.023	0.888	0.010	0.995	0.029	0.997	0.013
ACC	**T**	1.107	0.030	1.107	0.018	1.147	0.026	1.150	0.011
ACA	**T**	1.608	0.016	1.679	0.002	1.406	0.018	1.411	0.012
ACG	**T**	0.353	0.008	0.326	0.011	0.452	0.026	0.441	0.012
GCT	**A**	1.102	0.042	1.219	0.043	1.112	0.038	1.115	0.015
GCC	**A**	1.332	0.043	1.263	0.033	1.296	0.036	1.295	0.014
GCA	**A**	1.173	0.062	1.219	0.004	1.093	0.016	1.091	0.010
GCG	**A**	0.394	0.056	0.300	0.014	0.500	0.010	0.499	0.012

		East Africa		West Africa		Asia		America	
Codon	AA	Average	SD	Average	SD	Average	SD	Average	SD
TAT	Y	0.845	0.071	0.802	0.024	0.717	0.038	0.734	0.018
TAC	Y	1.155	0.071	1.198	0.024	1.283	0.038	1.266	0.018
CAT	H	0.933	0.053	1.066	0.011	0.818	0.037	0.809	0.015
CAC	H	1.067	0.053	0.934	0.011	1.182	0.037	1.191	0.015
CAA	Q	0.990	0.020	0.898	0.015	1.157	0.043	1.194	0.008
CAG	Q	1.010	0.020	1.102	0.015	0.843	0.043	0.806	0.008
AAT	N	0.741	0.061	0.795	0.022	0.644	0.018	0.652	0.020
AAC	N	1.259	0.061	1.205	0.022	1.356	0.018	1.348	0.020
AAA	K	0.813	0.014	0.776	0.005	0.870	0.023	0.873	0.007
AAG	K	1.187	0.014	1.224	0.005	1.130	0.023	1.127	0.007
GAT	D	0.809	0.021	0.817	0.019	0.951	0.034	0.935	0.014
GAC	D	1.191	0.021	1.183	0.019	1.049	0.034	1.065	0.014
GAA	E	0.956	0.014	0.880	0.009	0.930	0.014	0.928	0.006
GAG	E	1.044	0.014	1.120	0.009	1.070	0.014	1.072	0.006
TGT	C	1.075	0.026	1.020	0.085	0.923	0.067	0.888	0.026
TGC	C	0.925	0.026	0.980	0.085	1.077	0.067	1.112	0.026
CGT	R	0.387	0.012	0.409	0.022	0.447	0.058	0.472	0.013
CGC	R	0.671	0.014	0.664	0.014	0.601	0.069	0.572	0.016
CGA	R	0.319	0.031	0.338	0.031	0.230	0.020	0.218	0.010
CGG	R	0.530	0.047	0.523	0.026	0.554	0.039	0.576	0.015
AGA	R	2.581	0.030	2.448	0.013	2.393	0.020	2.398	0.013
AGG	R	1.513	0.031	1.619	0.014	1.775	0.037	1.764	0.014
GGT	G	0.524	0.028	0.550	0.008	0.513	0.016	0.520	0.012
GGC	G	0.658	0.019	0.636	0.024	0.692	0.012	0.691	0.011
GGA	G	1.958	0.031	1.915	0.014	1.732	0.024	1.735	0.011
GGG	G	0.860	0.037	0.898	0.003	1.063	0.028	1.054	0.012

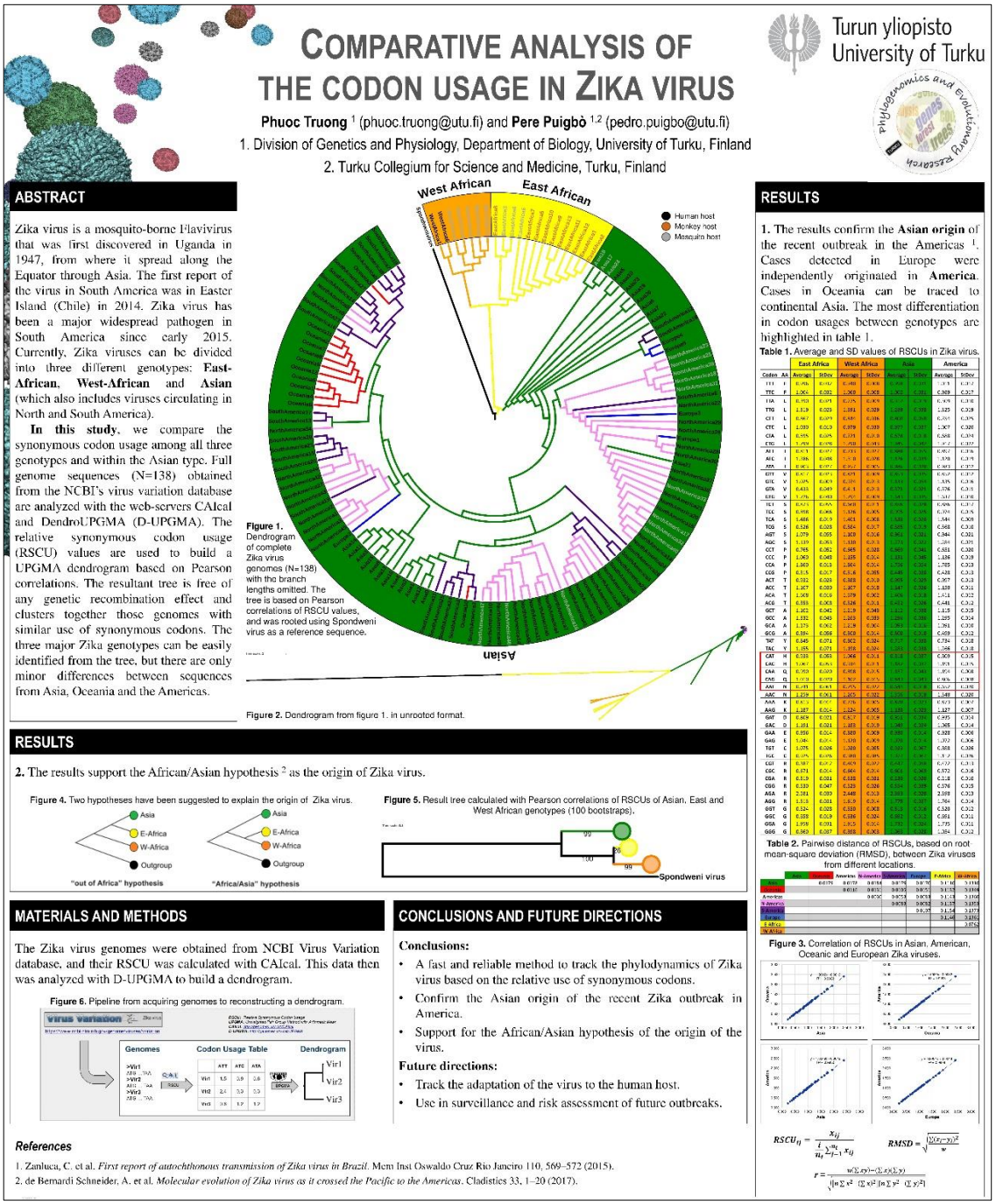

Figure 1. Poster of the Zika virus pilot study titled "Comparative analysis of the codon usage in Zika virus" by P. Truong and P. Puigbo. It was presented at OIKOS Conference for Ecologists and Evolutionary Biologists, which was hosted by the University of Helsinki and held in Helsinki (FIN) in January 2017.

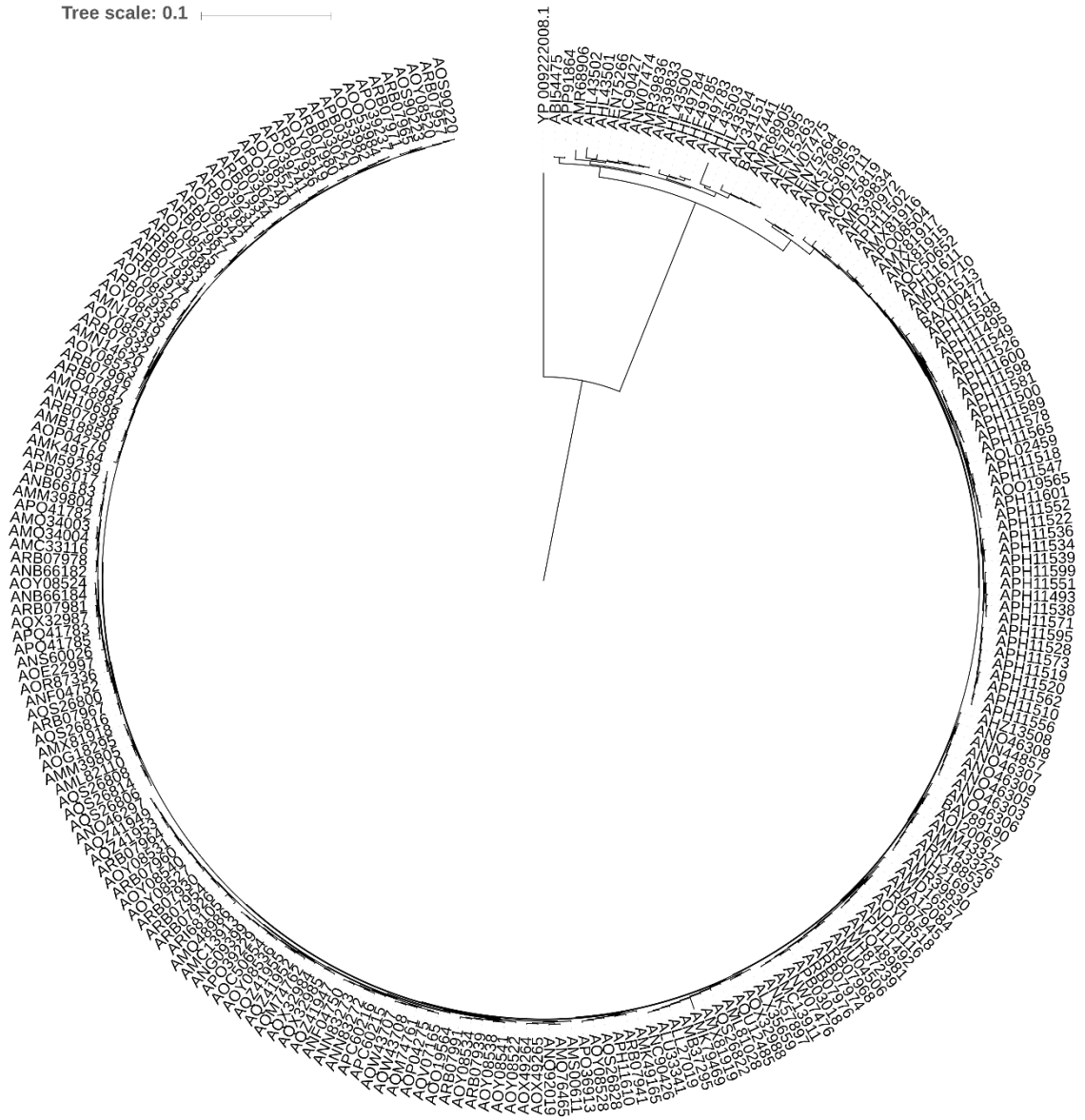

Figure 2. Phylogenetic tree of Zika viruses (N = 246). The tree was built by aligning protein sequences using MUSCLE and constructing it with FastTree with 100 bootstraps. The individual NCBI accession codes are shown at the tips of branches and the branch lengths are proportioned to Tree scale. Spondweni virus was used as an outgroup.

Appendix 3. Supplementary material related to optimal host identification

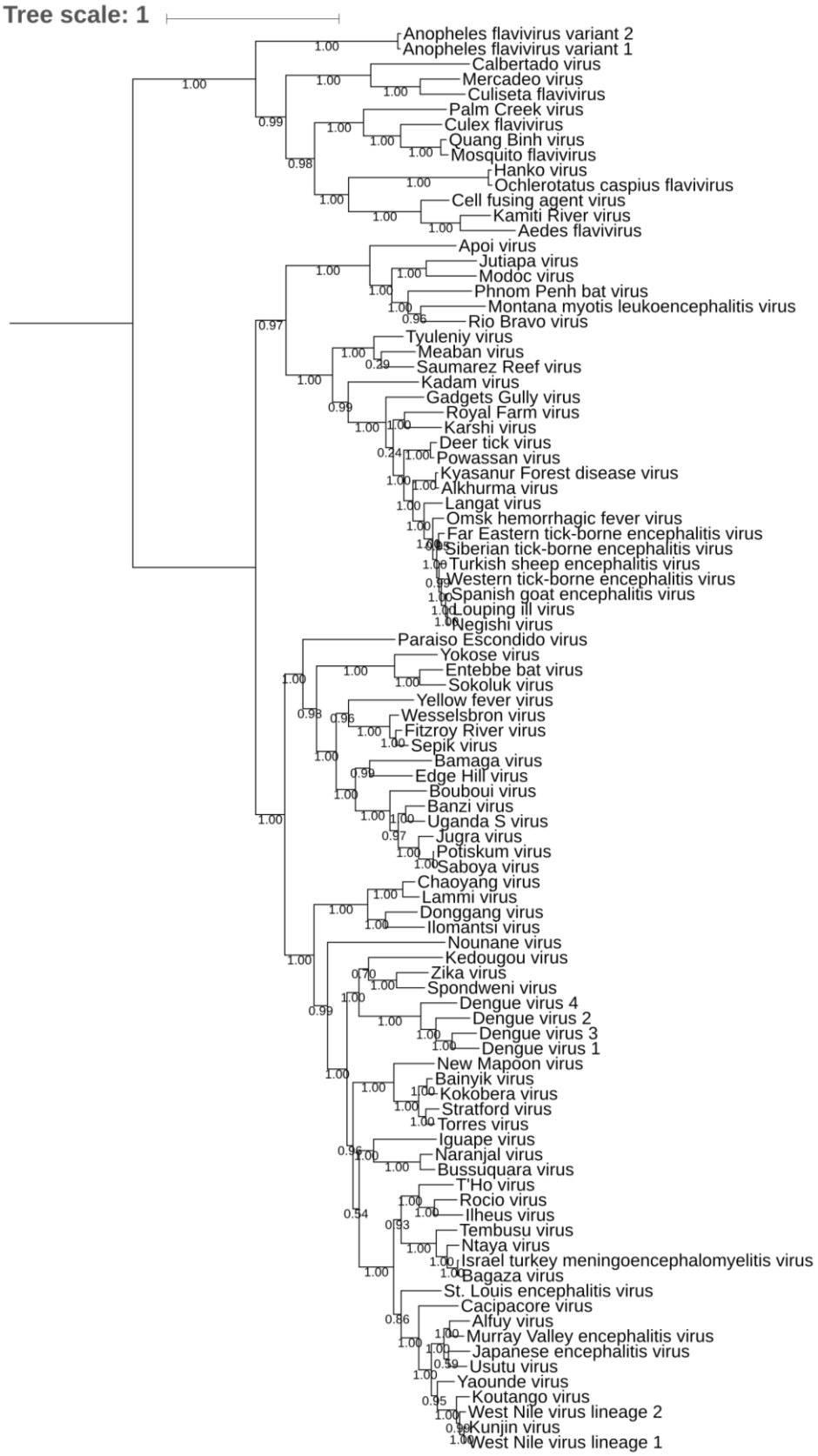

Figure 1. A comprehensive phylogenetic tree of flaviviruses (N = 94). The tree was computed from amino acid sequences with 100 bootstraps and rooted with the clade of insect-only flaviviruses. Branch lengths are proportioned to Tree scale.

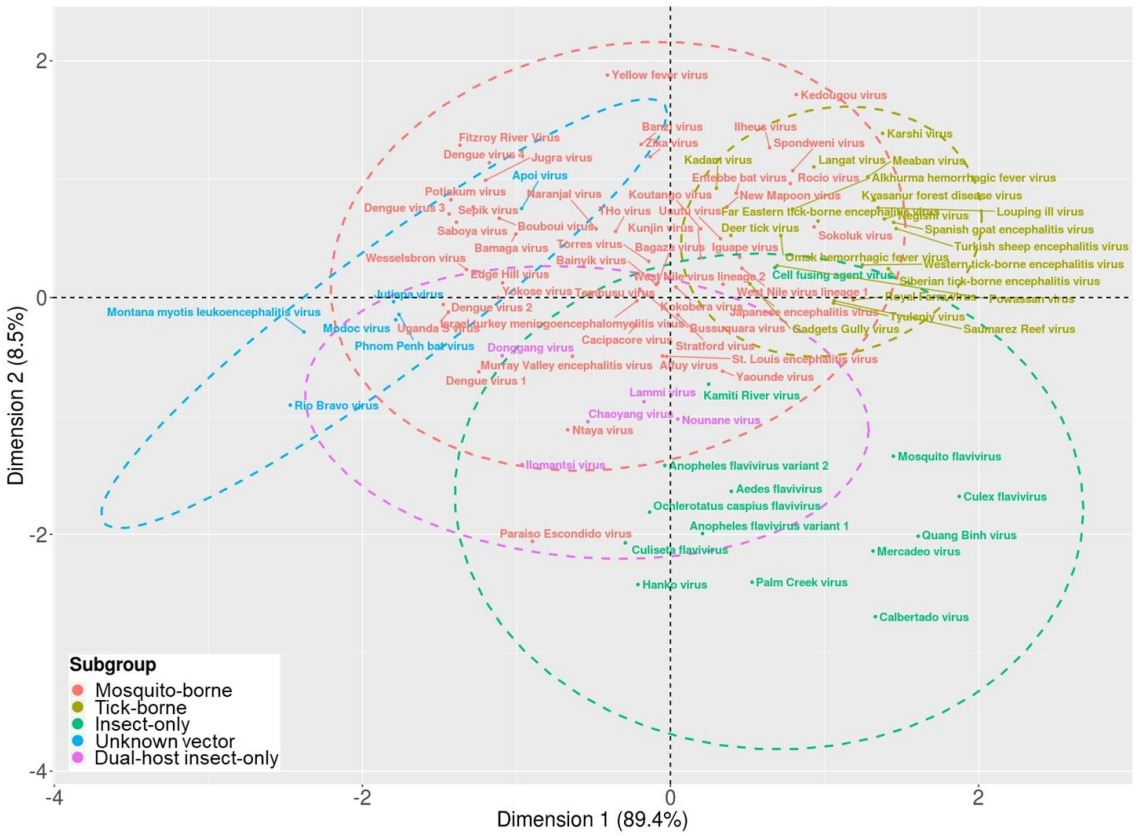

Figure 2. Interspecies correspondence analysis and subgroup centroids of normalized Codon Adaptation Index (nCAI) values of flaviviruses, genus *Flavivirus* (N = 94). While there is a distinct separation between the clusters of insect-only and the other flavivirus subgroups, the clouds (shown as dashed circles with colors corresponding to a flavivirus subgroup) computed from them show major overlap, indicating that most of the subgroups share similar host preferences. The centroids were computed based on the multivariate normal distribution of each subgroup with a confidence level of 0.95. Dimension 1 explains 89.4 percent and dimension 2 contributes to 8.5 percent of the variation.

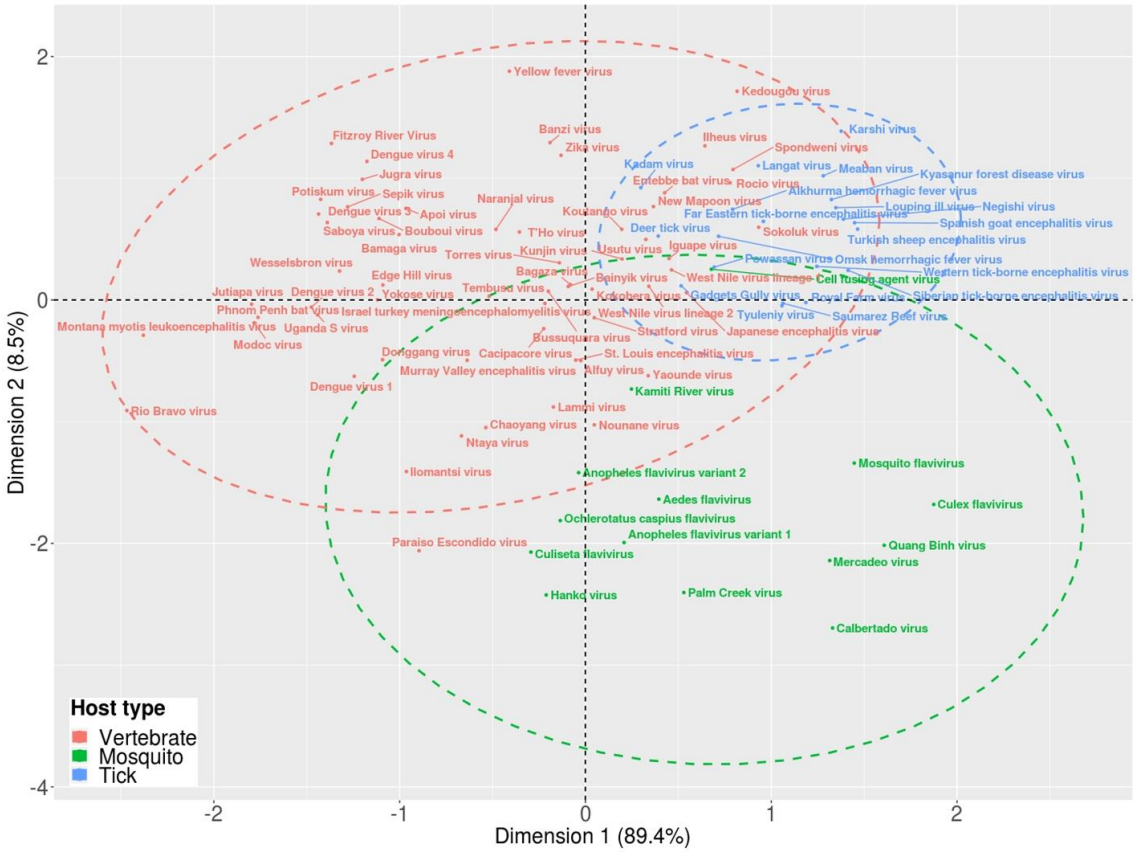

Figure 3. Interspecies correspondence analysis with host type centroids of normalized Codon Adaptation Index (nCAI) values of flaviviruses, genus *Flavivirus* (N = 94). When the flaviviruses are categorized based on their preferred host type, we can observe two major clouds (dashed circles) and clusters; viruses that have a vertebrate host and those that have a mosquito host. Tick-borne viruses largely overlap with the vertebrate host cluster, suggesting that while tick-borne flaviviruses cluster separately, they have the same type of host as mosquito-borne viruses. The centroids were computed based on the multivariate normal distribution of each host type with a confidence level of 0.95. Dimension 1 explains 89.4 percent and dimension 2 contributes to 8.5 percent of the variation.

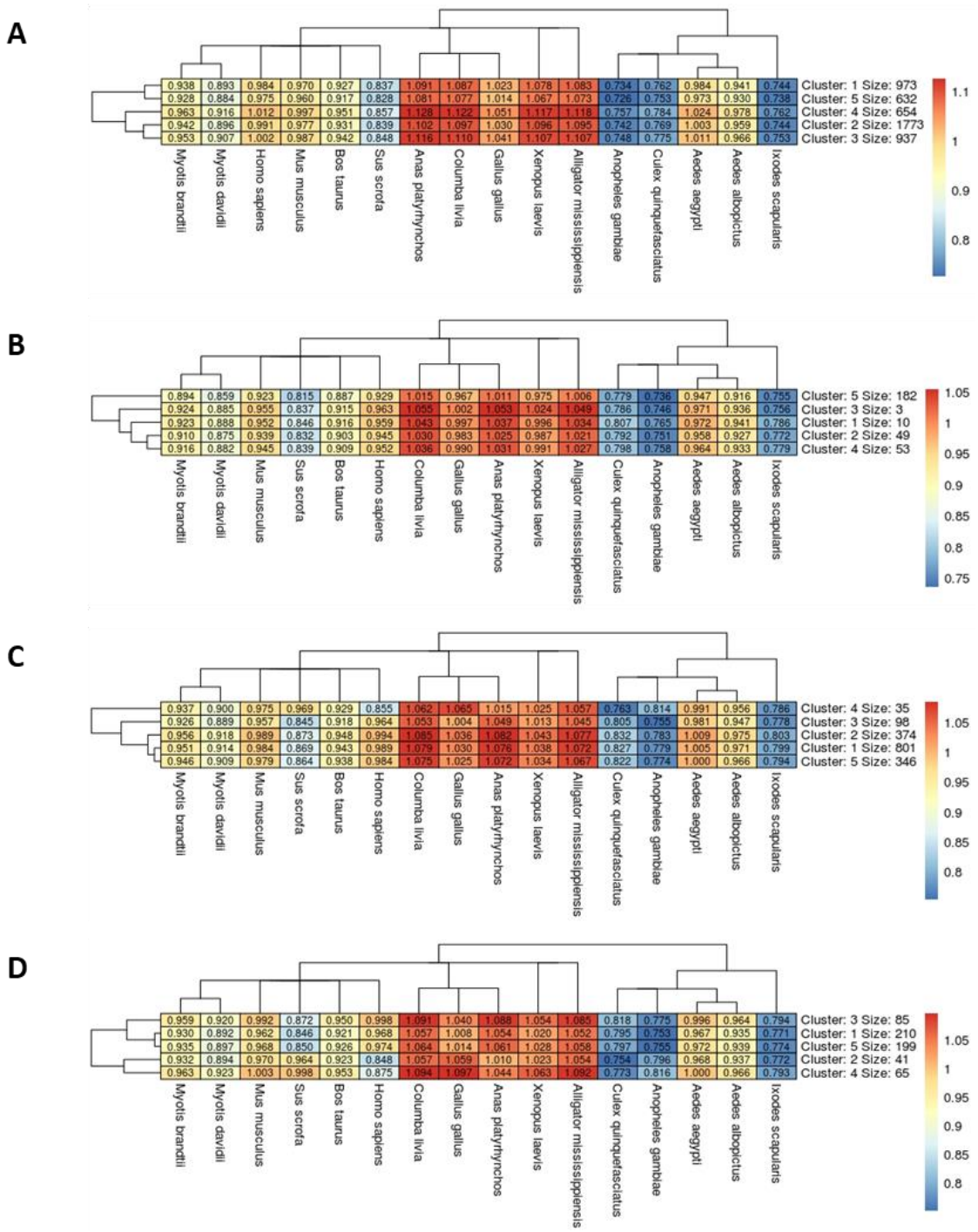

Figure 4. Intraspecies k-means (5) heat maps of normalized Codon Adaptation Index (nCAI) values of four major mosquito-borne flaviviruses (genus *Flavivirus*). When the nCAI values of (A) Dengue viruses, (B) Japanese encephalitis viruses, (C) West Nile viruses and (D) Zika viruses are plotted in heat maps, they display similar overall adaptation levels to different host organisms, although there are differences between these viruses. On average, the viruses are optimized for mice (*Mus musculus*), humans (*Homo sapiens*) and *Aedes* mosquitoes, especially *Ae. aegypti* (nCAI 0.95–1.05). The suboptimal hosts are the other mammalian hosts, avians, reptiles and amphibians due to overoptimization (nCAI > 1.05), and *Culex* and *Anopheles* mosquitoes, and deer ticks (*Ixodes scapularis*) due to underoptimization (nCAI < 0.95). The number and size of k-means clusters do not match the current classification of these flaviviruses.

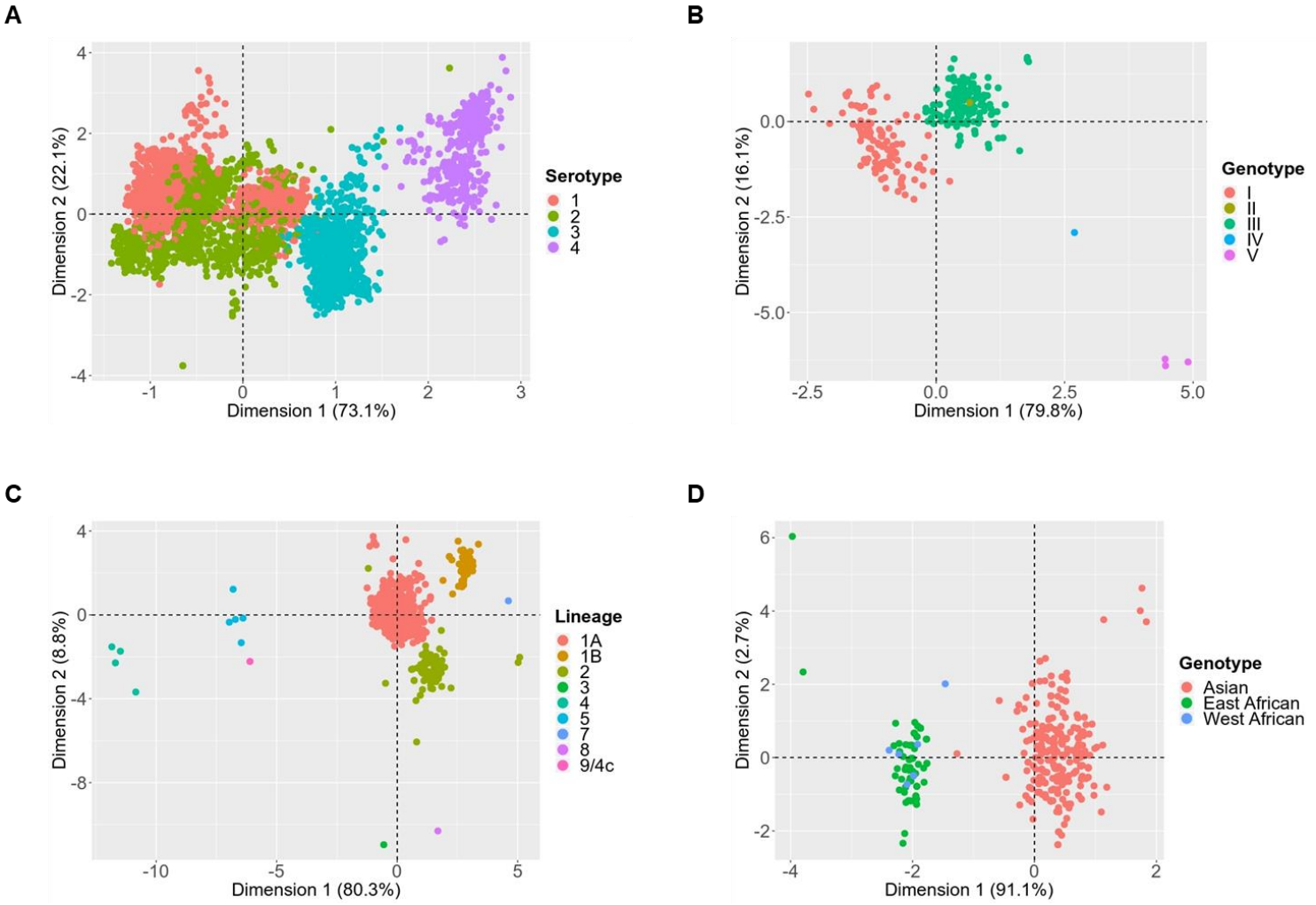

Figure 5. Intraspecies correspondence analyses of normalized Codon Adaptation Index (nCAI) values of four major mosquito-brone flaviviruses (genus *Flavivirus*). The results show that nCAI is able to discriminate between the different categories of (A) Dengue viruses (N = 4,865), (B) Japanese encephalitis viruses (N = 297), (C) West Nile viruses (N = 1,619) and (D) Zika viruses (N = 494). Additionally, the distances between separate clusters mirror actual phylogeny. In figure A, dimension 1 explains 73.1 percent and dimension 2 contributes to 22.1 percent of the variability. In figure B, dimension 1 explains 79.8 percent and dimension 2 contributes to 16.1 percent of the variability. In figure C, dimension 1 explains 80.3 percent and dimension 2 contributes to 8.8 percent of the variability. In figure D, dimension 1 explains 91.1 percent and dimension 2 contributes to 2.7 percent of the variability.

Appendix 4. Article in preparation.

AN UNSUPERVISED ALGORITHM FOR HOST IDENTIFICATION IN FLAVIVIRUSES

Phuoc Truong [1], Santiago Garcia-Vallve [2], Pere Puigbò [1,*]

[1] *Department of Biology, University of Turku, Turku, Finland.*

[2] *Department of Biochemistry and Biotechnology, Rovira i Virgili University, Tarragona, Catalonia, Spain*

[*] *Corresponding author: pepuav@utu.fi*

ABSTRACT

Surveillance and early characterization are essential to the control of emerging viruses. However, this remains challenging when new viruses emerge. A major challenge is the identification of the potential hosts of novel viruses. Here we introduce an unsupervised algorithm based on a normalization of the classical Codon Adaptation Index, which uses solely coding sequences, to identify the putative host-range in flaviviruses. The algorithm has been tested with genome sequences from 94 flaviviruses (including Dengue, West-Nile, Zika and several less common flaviviruses) and 16 potential hosts.