ARBOVIRUSES: THE HIDDEN PATH OF AN IMMINENT THREAT

by

Adriano de Bernardi Schneider

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2018

Approved by:

_____
Dr. Daniel Janies

_____
Dr. Jean-Claude Thill

_____
Dr. Jun-tao Guo

_____
Dr. Xinghua Shi

ABSTRACT

ADRIANO DE BERNARDI SCHNEIDER. Arboviruses: The hidden path of an imminent threat. (Under the direction of DR. DANIEL JANIES)

Arboviruses are a grade of viruses carried by arthropods, which have been in the headlines due to recent epidemics. Members of this grade are the families *Flaviviridae* which includes Zika (ZIKV), Dengue (DENV), Yellow Fever (YFV), among other viruses and *Togaviridae*, which includes Chikungunya (CHIKV).

Research on some arboviruses has been strong over the past couple of decades. Other arboviruses have not garnered much attention until lately. For example, ZIKV has been understudied until 2015. Since the 1950s ZIKV was considered to cause only a benign infection in humans. ZIKV became well studied only after the recent outbreaks of the virus in the Pacific, Americas, and South-East Asia, was found to be related to severe neuropathology, which includes the development of neurological defects such as microcephaly on the fetus and Guillain-Barré Syndrome in adults. CHIKV is another arbovirus that although been circulating for a long time in Africa and Asia, has been recently introduced into the Americas in 2013, causing recurring outbreaks in South and Central American naïve populations.

YFV, which been known to be endemic and thought to be controlled in South America, has re-emerged in Brazil beginning in December 2016. This outbreak, although restricted to transmission by the sylvatic mosquito *Haemagogus leococelaenus*, raised questions among researchers regarding the potential for spread to the United States due to the presence of the urban vectors *Aedes aegypti* and *Aedes albopictus* and a naïve, largely unvaccinated population. Another question that still remains is whether YFV will ever reach the Asian continent?

Today, the time it takes for awareness of the health organizations, to convince the funding agencies, and to work on vaccine development is much more than the

time needed for the disease to change from a local outbreak to a global epidemic. The overall objective of this work is to provide the grounds for a viral surveillance system based on evolution, utilizing the current ZIKV and CHIKV outbreaks and other arboviruses as case studies.

Utilizing phylogenetic and molecular sequence alignment tools I developed a pipeline to evaluate the genomic changes of viruses on CHIKV and ZIKV. I also created a pipeline to generate pathogen transmission networks and compare different disease networks utilizing different network centrality metrics. CHIKV, DENV, YFV and ZIKV were utilized as case studies. The strategies utilized in this work will enable better abatement and management strategies of viral outbreaks.

My findings indicate that changes in the coding sequence does not seem to be the main reason why ZIKV has changed its behavior in terms of pathogenicity. In CHIKV there is an insertion on the UTR region of a group of sequences and change of virulence has been associated with UTR sizes in different CHIKV strains. Upon analyzing viral 3' and 5' UTRs, a trinuncleotide motif, known as Musashi Binding Element was identified in both CHIKV and ZIKV, its presence and availability on ZIKV may explain a preference to human cells, in CHIKV the motif is present but not available. Although both CHIKV and ZIKV coexist and have spread in the same regions in a short period of time, their spread seems to be from independent events. When looking at transmission networks, there is a high correlation between the different centrality metrics utilized to measure all four DENV serotypes transmission networks, CHIKV, YFV and ZIKV have lower correlation, thus, distinct patterns.

ACKNOWLEDGEMENTS

I would like to thank Dr. Daniel Janies for the immense support during my period at UNC Charlotte. Also, the advisory committee Dr. Jean-Claude Thill, Dr. Jun-tao Guo and Dr. Xhinghua Shi.

Dr. Michael T. Wolfinger, research scientist at the Department of Theoretical Chemistry, University of Vienna for the insights on the structure of RNAs, Dr. Tzu-Hao Chang, Associate Professor at the Graduate Institute of Biomedical Informatics from Taipei Medical University, Taipei, Taiwan for assisting on mapping RNA motifs with RegRNA 2.0, Dr. Denis Jacob Machado, Department of Zoology, Institute of Biosciences, University of São Paulo for the great phylogenetics discussions, Michael Cioce and John Williams for their support on building the transmission networks and my intern, now Doctor-to-be, Lambodhar Damodaran for always performing above of the expected.

Would also like to acknowledge the Graduate Assistant Support Plan (GASP) from UNC Charlotte and the Department of Bioinformatics and Genomics for the tuition and graduate assistantship during the course of my doctoral studies.

DEDICATION

For Sofia Darrieux Schneider, my guiding star.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CHIKV An acronym for Chikungunya virus.

DENV An acronym for Dengue virus.

EA-IOL An acronym for Chikungunya Eastern Africa - Indian Ocean Lineage

ECSA An acronym for Chikungunya East, Central and South African Lineage

GBS An acronym for Guillain-Barré Syndrome.

IOL An acronym for Chikungunya Indian Ocean Lineage

MA-SA An acronym for Chikungunya Middle Africa - South America Lineage

MAYV An acronym for Mayaro virus.

MBE An acronym for Musashi Binding Element.

ONNV An acronym for O'nyong nyong virus

POWV An acronym for Powassan virus.

SHR An acronym for Source Hub Ratio

UTR An acronym for Untranslated Region

WHO An acronym for World Health Organization

WNV An acronym for West Nile virus.

YFV An acronym for Yellow Fever virus.

ZIKV An acronym for Zika virus.

CHAPTER 1: INTRODUCTION AND BACKGROUND

Researchers have been describing recent outbreaks of arboviruses that have not been previously detected for decades ([3, 4]). The issue of emerging diseases is not news for viruses in the broad sense, but these have a peculiarity, they are all transmitted by the same group of vectors: mosquitoes and ticks, and belong to groups alike, *Togaviridae* and *Flaviviridae*.

The recent outbreaks have brought concern due to the lack of basic knowledge surrounding the disease caused by them. Chikungunya virus been first isolated in Tanzania in 1952 [5], and, despite some travel-related cases, has been only found in the Americas as an autochthonous case on October 2013 on the island of Saint Martin [6]. Zika virus (ZIKV), known as a mild disease isolated in the Ziika forest in Uganda in 1947 [7], had its first cases in the Americas on March 2014 in Easter Island (Chile), to later in 2015 be confirmed that there were clusters of cases in Brazil on February 2015 [8].

For the purpose of this work, I selected Chikungunya virus (CHIKV) and ZIKV, as they are carried by mosquitoes and belong to a non evolutionary group called "arboviruses", as they belong to different taxonomic groups. ZIKV is a *Flavivirus* belonging to *Flaviviridae*, while CHIKV is an *Alphavirus* that belongs to *Togaviridae*. Dengue virus (DENV) was not selected for this analysis given the time of introduction in the Americas diverge from ZIKV and CHIKV by hundreds of years. ZIKV and CHIKV are more comparable due to recent introduction to the Americas. The transmission network of DENV, along with Yellow Fever virus (YFV) will be discussed on the Chapter 3 of this dissertation.

1.1 Chikungunya disease

Chikungunya name comes from the Makonde language, spoken by an ethnic group in southeast Tanzania, and means the bent posture that persons with severe arthralgia (joint pain) stay at when suffering of Chikungunya fever [9]. CHIKV is an *Alphavirus* close related to O'nyong nyong virus (ONNV), a virus previously considered a subtype of CHIKV. CHIKV consists of a positive-sense single-stranded RNA molecule of approximately 11.8kb, encoding nine proteins, which are divided in structural and non-structural proteins [10]. The disease have not been associated with life threatening symptoms until a recent outbreak the Réunion island, an overseas department from France located in the Indian Ocean, east of Madagascar, in 2005 [11].

CHIKV disease has been usually associated to fever, which lasts about 1 week; myalgia, which lasts between 7-10 days; polyalthralgia and polyarthritis, which can last from weeks to months; and rashes, which last about 1 week. Viremia generally lasts 5-7 days. After the outbreak in Réunion, higher morbidity has been observed, as well as neurological issues, such visual and hearing loss, paralysis and Guillain Barré Syndrome (GBS) and renal complications [12, 13, 14].

1.2 The geographic spread of Chikungunya virus

First isolated in the Newala district of Tanganyika, a British territory that is currently part of Tanzania [15, 16], CHIKV today can be considered a virus with at least four lineages: Asia lineage, Indian Ocean Lineage (IOL), East, Central and South African Lineage (ECSA), and West African [17]. Its origin and the history of how it spread inside Africa remains to be discovered, given the lack of sampling and presence of confounding events such as the presence of co-circulating DENV and YFV, although its presence in sylvatic mosquitoes and nonhuman primates in Uganda and Tanzania point that they likely originated in Central/East Africa [18]. The emergence and spread is believed to have started as early as the 18th Century through sailing

ships, as stored water facilitated mosquito reproduction [19].

It is known that the ECSA lineage gave rise to the Asian and IOL from independent outbreaks that are several years apart. The Asian lineage is predicted to have emerged from ECSA somewhere between 50 to 300 years ago, it caused outbreaks in Southeastern Asia and India, being present in Southeastern Asia until today [20, 21]. The IOL has emerged from another outbreak from a different strain from ECSA which started to spread in 2004, creating the Kenya epidemic, followed by Southeastern Islands of the Indian Ocean in 2005 and Réunion Island in 2006 where 40% of population was seropositive [22, 23]. Since then, IOL emerged in outbreaks in India in 2005 and Southeastern Asia countries in 2008, as well as two outbreaks in Europe in 2007 and 2010. IOL is currently present in India and is co-circulating with the Asian strain in Southeastern Asia. On this specific lineage, it has been observed a mutation that increased the competency of *Aedes albopictus* as vector, increasing the infection rate and extending the range of the disease [24].

The West African lineage is described in the literature as a sister clade to the ECSA, IOL and Asian lineages and curiously, have been restricted to only West Africa. On the other hand, the Asian strain has emerged in Central America in 2013 and has been found in South America as well [25]. Also, in 2014 there was an ECSA lineage outbreak in Brazil.

The first emergence in urban cycle is believed to be between 1879 and 1956, when the enzootic lineage ECSA was introduced into Asia. This introduction created the Asian lineage, which caused outbreaks in India and Southeast Asia and still circulates in Southeast Asia.

1.3 Zika virus disease

ZIKV belongs to the *Flaviviridae* family, which consists of positive sense single stranded RNA viruses such as DENV, YFV, West Nile virus (WNV), and Hepatitis C virus among others. The ZIKV reference sequence genome (National Center for

Biotechnology Information (NCBI ncbi.nlm.nih.gov) accession NC 012532.1 strain MR 766 from Uganda) comprises 10,794 base pairs of linear RNA. The genome has 5' and 3' untranslated regions that form complex RNA structures which interact and regulate the replication, protein expression and tissue tropism of the virus while it infects a host cell. The remainder of the ZIKV genome codes for a 3419 amino acid polyprotein that encodes at least 12 proteins [26].

The virus was named after the place where it was first discovered and isolated, the Ziika forest in Uganda in 1947 [7]. At the time, it was isolated from rhesus monkeys and the disease was considered mild with symptoms being only fever, body pain and skin rashes. It was then identified in humans in Africa as well as in Malaysia in 1966 and there are serological reports of the presence of the virus in Southeast Asia (Indonesia, Cambodia and Malaysia) and Pakistan since the 1960s [27, 28]. But it was not until 2015 that certain research groups in South America and French Polynesia realized that ZIKV could be related to the GBS and microcephaly [29, 30, 31]. In the same year, it was observed that the virus was actually causing a fetal neuropathogenesis, which had as its most evident symptom the microcephaly, but other fetal abnormalities such as loss of hearing and vision were related [32].

A more severe ZIKV disease spectrum (GBS and fetal syndrome) was observed in both the 2013-2014 outbreak in French Polynesia and in the Americas [33] [34]. Recently, viral sequences representing the Asia-Pacific-Americas clade have been recovered from both primary microcephaly-associated central nervous tissue and amniotic fluid from affected pregnancies [35, 36].

Comparison of the sample of viral genetic sequences available today shows that mutations have occurred as ZIKV has spread from Asia across the Pacific to the Americas. However, there is a 41-year sampling gap between the Malaysian isolate from 1966 and the isolates from Yap in Micronesia in 2007. Disease oriented questions include whether the change in ZIKV disease phenotype correlates to changes in ZIKV

viral genotype, or if the current disease spectrum is influenced by the virus spreading within a host population with particular susceptibility or confounding effects. These effects can include the genetic background of humans and/or previous infection with another virus such as DENV or environmental effects [37].

1.4 The geographic spread of Zika virus

As mentioned above, the first sequence isolated for the ZIKV is from a sentinel rhesus monkey from the Ziika Forest, a tropical forest near Entebbe, Uganda, where researchers from the Yellow Fever Research Institute carried surveillance work [38]. Contemporaneous to identifying ZIKV seropositivity in monkeys and subsequently in human populations in Africa, similar evidence suggesting human infection was collected in Egypt [39], India [40], Malaysia [41], Thailand [41], Vietnam [42] and the Philippines [43]. Based on this serologic history, ZIKV may have circulated in parallel in Africa and Asia during the first half of the 20th century or earlier. Weaver et al., 2016 [44] caution that the serology surveys may have been conducted with reagents that are cross-reactive among flaviviruses, generating false positives. No sequence data is available to the public from Asian and Egyptian isolates from this time frame except for a Malaysian isolate from 1966 from an *Aedes spp.* host (NCBI accession HQ234499).

Prior sequence analysis has described ZIKV as being comprised of three clades: a West African (Nigerian cluster), East African (strain name MR766 cluster), and an Asian clade reconstructed by some to have originated from East Africa [45]. The ZIKV lineage that has been infecting patients in the Americas since 2014 [46] is genetically similar to ZIKV previously isolated from Micronesia [47], French Polynesia [48], and Easter Island [49]. While the ZIKV in South America, Central America, the Caribbean, and Mexico are descendants from the Asian lineage, the ZIKV in these regions is now considered a novel Asia-Pacific-Americas lineage [50]. This lineage continues to spread across the globe, leading to recent outbreaks occurred in Cape

Verde and Bali [51].

The occurrence of ZIKV outside of Africa has been the focus of few studies until 2015. Epidemiologists investigating the 2007 ZIKV outbreak in the four islands that comprise Yap, in the Federated States of Micronesia, did not detect severe clinical outcomes [47]. Since 2007 there have been cases of ZIKV in the western Pacific including: the Philippines and Thailand. In 2014, ZIKV had reached Easter Island [4] and Haiti [52]. These discoveries contradict reports that ZIKV entered the Americas first in Brazil [53]. ZIKV was not reported in Brazil until early 2015 [46].

CHAPTER 2: THE PATH OF TWO ARBOVIRUSES

2.1 Introduction

Two arboviruses, CHIKV and ZIKV, from different families but with similar range of vectors have re-emerged in the past decade in a similar manner and increased pathogenicity. Although multiple explanations have been proposed to elucidate the recent changes in pathogenicity, the actual mechanism by which they changed remains unclear, as well as the reason of why they re-emerged.

By using phylogenetic and geographical analyses I investigate the emergence and spread of mutations and structural differences in CHIKV and ZIKV viral genomes that are correlated with changes in diseases associated with both viruses. I also mapped characters of interest, such as metadata (location, date, host) and apomorphies (mutations unique to specific clades/groups), to phylogenetic tree in order to look for related changes in disease behavior with viral mutations. I evaluate viral 3' and 5' untranslated terminal RNA sequences (UTRs) as they are related to viral replication and gene expression and examine the presence of regulatory elements and the effect *in silico* of mutations in elements of interest among different sequences of the viruses. By mapping such mutations and evaluating UTRs *in silico*, we are able to have good pointers to what is actually happening *in vivo* at the same time we avoid the expense and danger of gain of function experiments [54].

2.2 Material & Methods

All the geographic, sequence and temporal metadata for ZIKV and CHIKV available in the public domain (NCBI) as of April 1st, 2018 was consolidated in order to create datasets for the analyses described below (Figure 2.1). Selection of the appro-

priate outgroup for ZIKV and CHIKV was be based on the phylogeny of flaviviruses and alphaviruses, respectively. To investigate the phylogenetic relationships, I used the maximum likelihood tree search method as implemented in IQ-TREE [55]. I also used the Recombination Detection Program (RDP) [56] to investigate the possibility of genetic recombination among lineages.

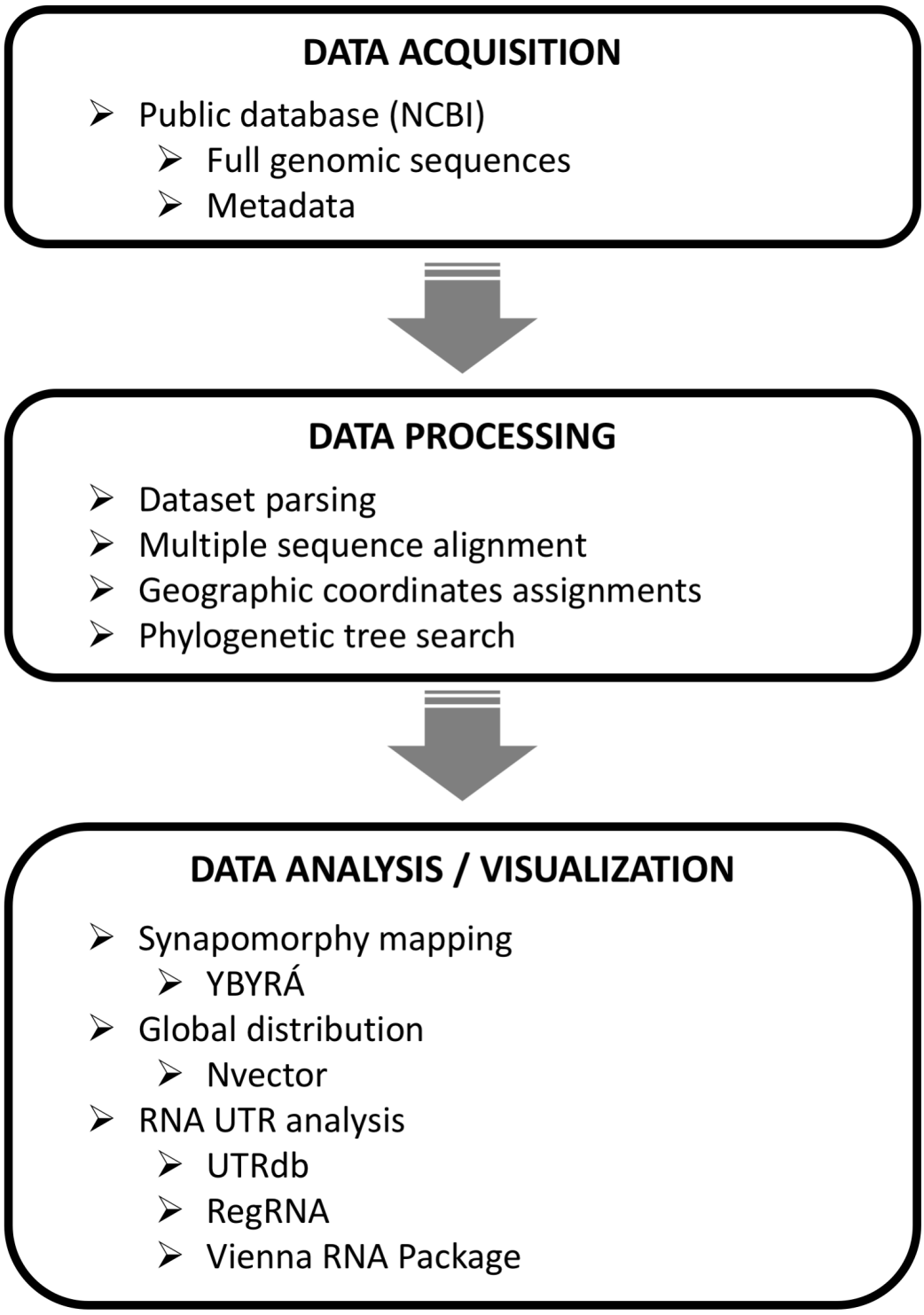

Figure 2.1: Data analyses Pipeline.

2.2.1 Multiple sequence alignment

Multiple sequence alignment for all datasets were be aligned using MAFFT v7.215 [57] under default settings. The alignments were visualized in Geneious [58]. Ragged edges resulting from differences in laboratory finishing may have been trimmed and marked as missing data.

2.2.2 Datasets

Three nucleotide datasets were created for ZIKV and CHIKV virus: 1) One genome dataset with partial or complete 5' and 3' UTR in addition to the polyprotein with an outgroup sequence. 2) One dataset that include all available 3' UTRs 3) One dataset that included all available 5' UTRs. Dataset 1 was subdivided in smaller datasets according to the major clades for the analysis of global distribution of the virus utilizing NVector (described below).

2.2.3 Outgroup search

To determine the appropriate outgroup for CHIKV, the closest related virus based on previous phylogeny of the *Alphavirus* was selected and the full genomic sequence was aligned to the ingroup taxa using MAFFT [57]. Given the single directionality on the previous trees of Zika virus and not novel sequences in Africa that could create a bias in topology due to long distance on the outgroup branch, the oldest African isolate was selected as the outgroup.

2.2.4 Phylogenetic tree search

Phylogenetic tree search was conducted on the genome dataset for both viruses using IQ-Tree [55] utilizing the implemented substitution model. IQ-Tree also examines the utility of other (GTR+G and GTR+I+G) and mixed models of nucleotide substitution for the following partitions within the virus genomes based on their genome structure (proteins within polyprotein).

2.2.5 Place of isolation metadata analysis

CHIKV and ZIKV phylogenies were visualized and character reconstruction for place of isolation metadata under parsimony was conducted in Mesquite v 3.04 [59]. Final trees with metadata were rendered utilizing FigTree v 1.4.2 [60]. I projected the phylogenetic trees into a virtual globe with a program developed in-house named NVector (described below).

2.2.6 Global distribution of virus phylogenetic tree on Nvector

CHIKV and ZIKV sequence and metadata containing location information were parsed and geographic coordinates for each sequence were obtained using LatLong.net. These data, along with isolation date for each sequence were combined with the phylogenetic tree data and plotted into the world map using the in-house program Nvector [61].

2.2.7 Synapomorphy mapping

Sequence alignments of protein and/or nucleotide were mapped into their proper phylogenetic tree to identify synapomorphies (derived changes shared by descendants) utilizing YBYRÁ [62].

2.2.8 RNA UTRs analyses

The 5' and 3' UTRs of ZIKV and CHIKV were annotated using the UTRScan tool from the UTRdb [63] and RegRNA 2.0 [64] and elements identified were reviewed individually according to possible relevance to change in pathogenicity.

The Musashi Binding Element (MBE), a trinucleotide motif, was selected for further investigation given that previous literature indicated that MBE was involved in increased viral replication of ZIKV and congenital defects in humans [65, 32]. A biophysical model was employed at the level of secondary structure and the opening energy of trinucleotides in the ZIKV and CHIKV UTR's on a shuffled sequence context and z score statistics were calculated. The opening energy of a region within

an RNA sequence is directly related to the local RNA secondary structure. According to the previous statement, the low opening energy is an indicator for single-strandedness, which correlates to the accessibility of the motif to bind to the protein. This approach is implemented in the Perl utility plfoldz.pl, which is available from https://github.com/mtw/plfoldz. The script employs the ViennaRNA [66] scripting language interface for thermodynamics calculations, the ViennaNGS [67] suite for extraction of genomic loci and the uShuffle Perl bindings [68] for k-let shuffling. The tool reports for each requested trinucleotide the opening energy in a genomic context as well as an opening energy z score obtained from n shuffling events of upstream and downstream sequences. Here, $n = 10,000$ dinucleotide shuffling events were used.

For ZIKV, the analysis of the relative binding energy MBE to the Musashi protein was performed using methods described by Zearfoss et al. [1]. The previously identified and characterized twelve base sequence including four bases 5' and five bases 3' of the conserved UAG MSI core motif were analyzed for representative ZIKV isolates [2]. Alignment comparisons of sequences to identify sequence relationships between available full sequence isolates were performed using MAFFT [57] and visualized on Geneious [58].

2.3 Results

2.3.1 Rooting and outgroup choice

Most studies tend to utilize midpoint rooting as the rooting strategy, not giving directionality to the tree. Although on one side it avoids the selection of a wrong outgroup that can create a false directionality within the tree, midpoint seems an arbitrary choice, whereas outgroup is intentional and can be backed up by the fact that a good choice will join the ingroup and outgroup at a point where a deep common ancestor logically would be found. Its been seen, for example, that midpoint rooting misled the search for the zoonotic origins of SARS-CoV [69].

A previous study on ZIKV evolution has shown that there is no change of tree

topology when selecting the oldest sequence MR766 (LC002520) or the closest related virus, Spondweni [2]. When selecting the oldest sequence as outgroup on CHIKV I observed a complete different topology from midpoint rooting and previous literature. In order to make sure the topology was correct and utilize the outgroup criterion, I generated an alignment with the closest virus related to CHIKV, ONNV, which caused the topology to agreed with previous studies and conserved the major clade relationships. Different from ZIKV, in the case of CHIKV, the presence of multiple old isolates from distinct clades makes difficult to select an outgroup within CHIKV to satisfy the outgroup criterion.

2.3.2 Phylogenetic Trees

The phylogenetic tree search performed with IQ-Tree resulted in two optimal Maximum -likelihood trees, one tree for CHIKV with 697 genomic sequences, including ONNV (NC_001512.1) outgroup sequence (Figure 2.2). The phylogenetic tree generated for ZIKV contains 491 genomic sequences with (LC002520) as the outgroup sequence (Figure 2.3).

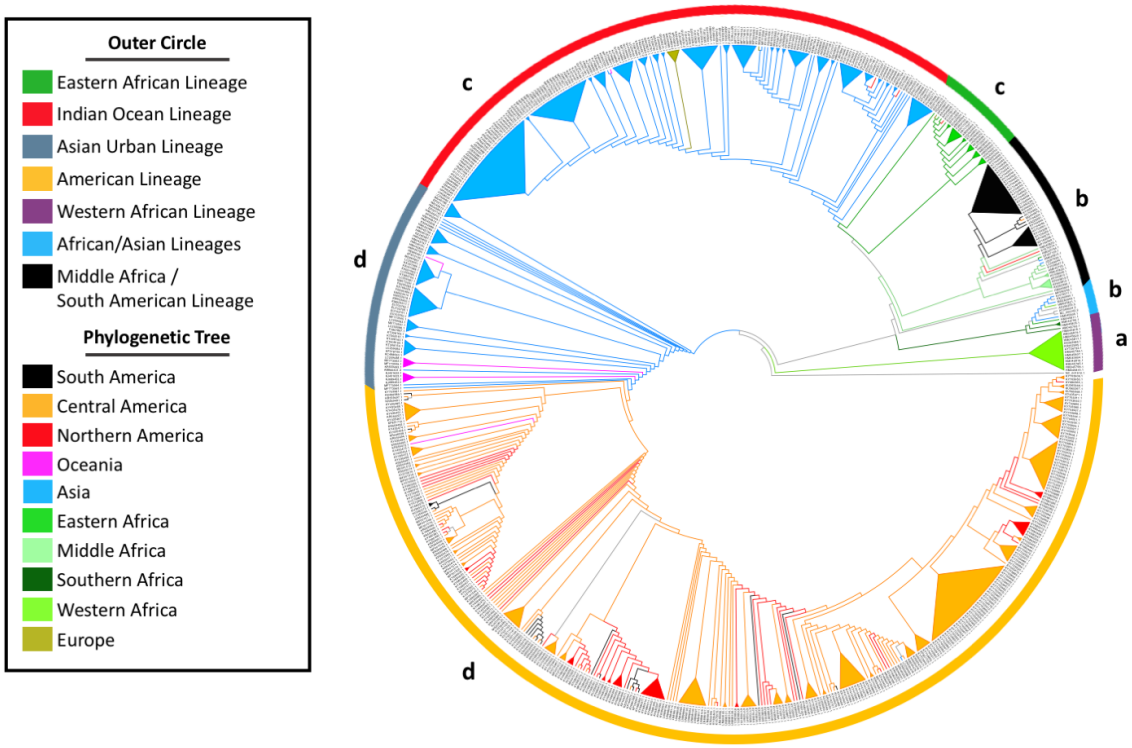

Figure 2.2: Maximum-Likelihood phylogenetic tree of 697 Chikungunya virus genomic sequences. Outgroup = O'nyong nyong virus - NC_001512.1. a = West African lineage (Figure 2.4) / b = Middle Africa/South America lineage (Figure 2.5) / c = Eastern Africa / Indian Ocean Lineage (Figure 2.6) / d = Asian Urban Lineage (Figure 2.10).

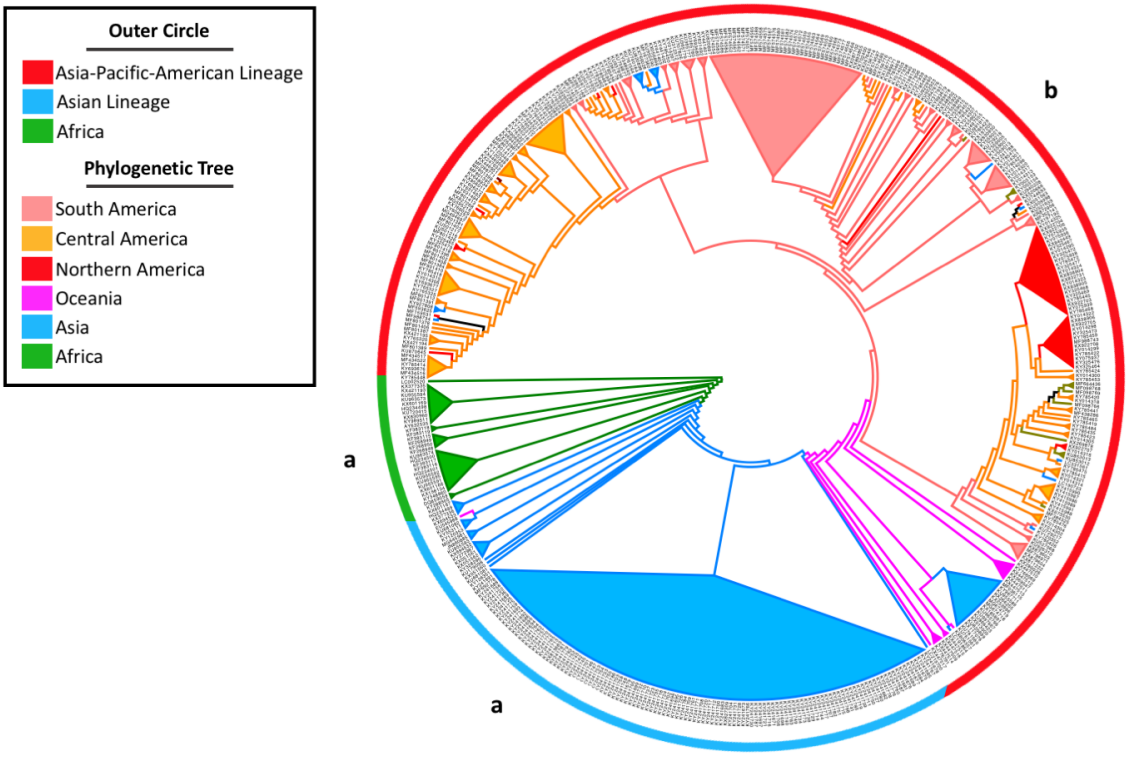

Figure 2.3: Maximum-Likelihood phylogenetic tree of 491 Zika virus full genomic sequences. a = African strains and Asian lineage (Figure 2.8) / b = Asia Pacific American lineage (Figure 2.13). Note: Large Asian Lineage clade is due to a large sequencing effort of ZIKV patients during a short period of time in the Singapore outbreak.

2.3.2.1 Chikungunya virus

The strains grouped according to known literature on the phylogenetic tree generated for CHIKV (Figure 2.2), which was split into subtrees in order to better understand the relationship between the taxa for each major clade as well as to map the synapomorphies found using YBYRÁ [62]. Three main clades can be observed for CHIKV, which can be divided in four subtrees, labeled as follows based on geographic spread:

1. EA-IOL lineage - Eastern African strains form a monophyletic group with Asian strains known as IOL (Figure 2.6).

2. MA-SA lineage - A sister clade to Eastern African strains, Middle African strains

form a monophyletic group with South American strains creating a new lineage that moved from Middle Africa to South America (Figure 2.5).

3. Asian Urban lineage - In a sister clade to Eastern and Middle African strains, Asian sequences known in the literature as Asian Urban lineage formed a clade that has strains from Oceania and the recent introduction of the virus in Central and Northern America (Figure 2.7).

4. West African lineage - Western African strains group as a sister clade to all other strains (Figure 2.4).

For the purpose of this work, I split the ECSA lineage into 3 regions, East Central and South Africa. As seen on the phylogenetic tree, the known ECSA/IOL is treated as a single lineage that started in Eastern Africa and moved to Asia (EA-IOL), the ECSA/South America is treated as a single lineage that started in Middle Africa and moved to South America (MA-SA). Although the Southern African lineage is related to a few Asian isolates, no major clade has been formed from outbreaks originating from that geographic region in Africa.

Due to the elevated number of strains with missing protein annotations on NCBI for CHIKV, I chose to evaluate only at the nucleotide level to avoid bias on the results due to missing data. The West African lineage clade had the largest number of non-ambiguous synapomorphies, 640. The Asian Urban lineage clade had the second largest number of non-ambiguous synapomorphies, 371. The American strains within the Asian Urban lineage clade had 5 non-ambiguous synapomorphies, with a section of the subtree (noted on Figure 2.7) with 30 non-ambiguous synapomorphies (insertions) occurring all on the UTRs of the CHIKV genome. The EA-IOL clade had 113. The MA-SA lineage have 14 synapomorphies shared among Middle Africa and South America, and 139 exclusive to South American strains. The sister clade to MA-SA lineage, which encompass sequences from Asia, Middle and Southern Africa,

had 50 synapomorphies. Only synapomorphies on key nodes (new clade/lineage based on geography) were evaluated.

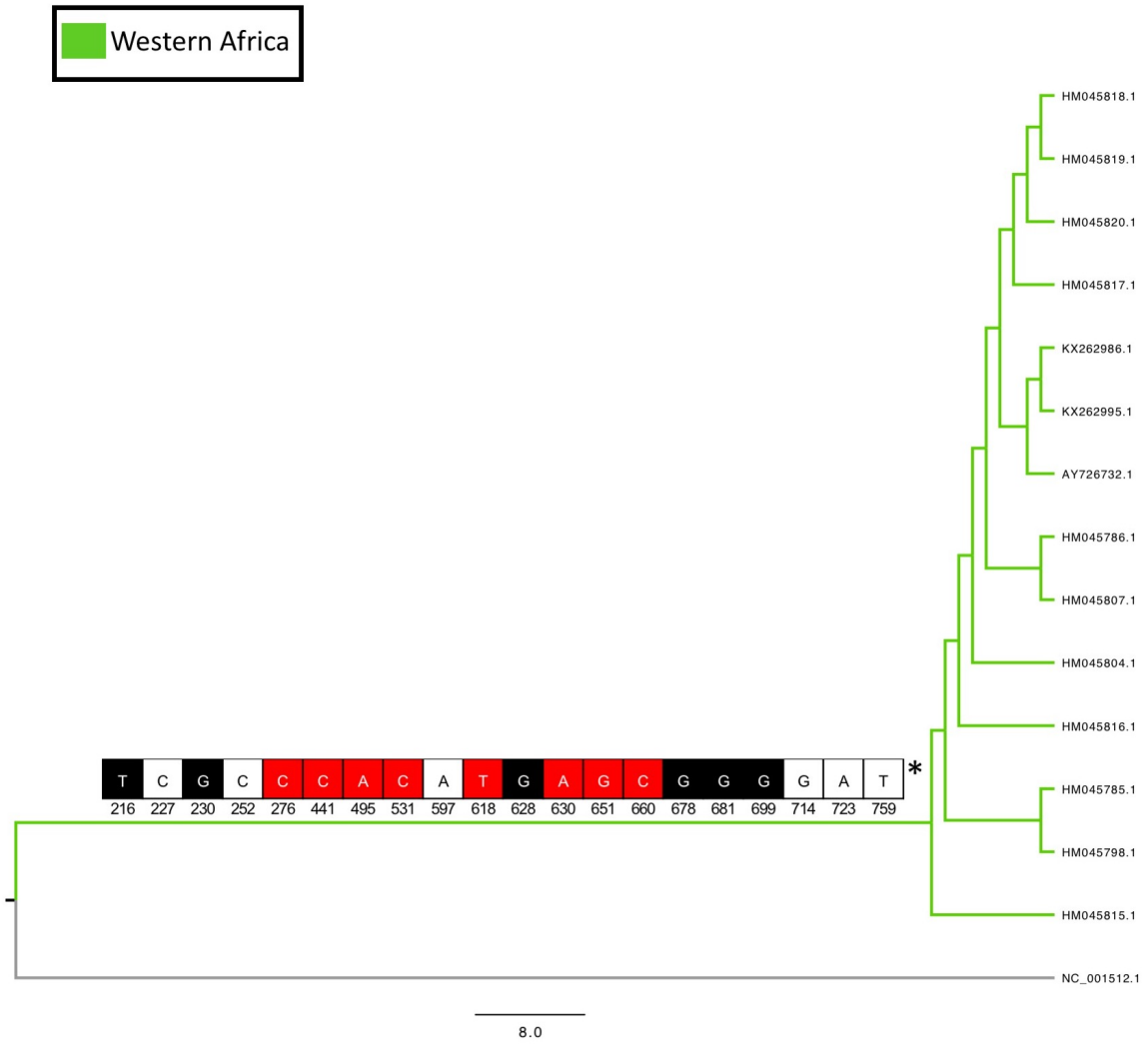

Figure 2.4: Chikungunya virus subtree - West African lineage clade - unique nucleic synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, White cells are ambiguous optimized characters. Node 715 (marked with star) first 20 synapomorphies shown, all other can be found on Appendix Table B.1.

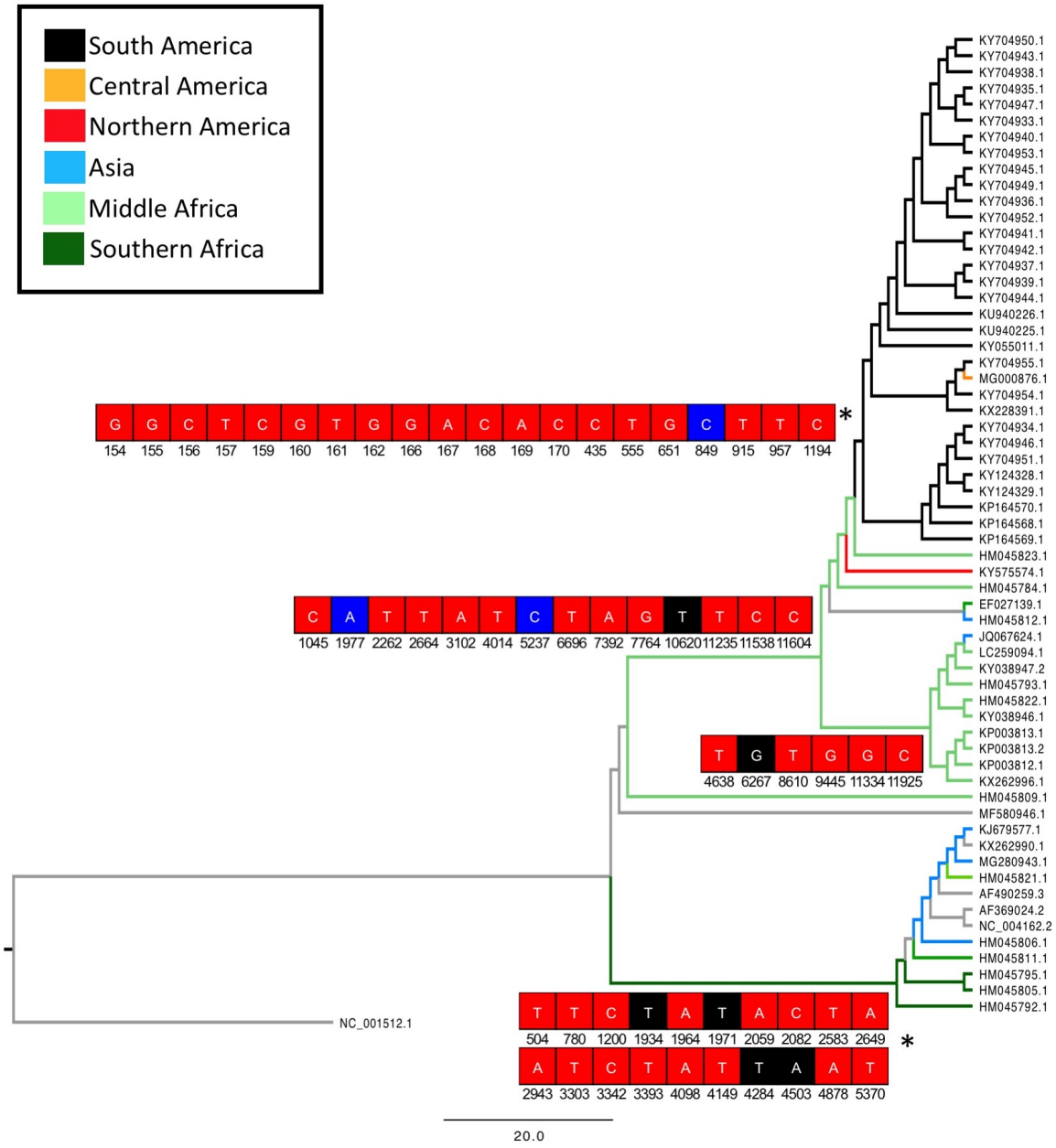

Figure 2.5: Chikungunya virus subtree - Middle Africa/South America lineage clade - unique nucleic synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, Blue cells are non-unique, homoplastic synapomorphies. Nodes 928 and 705 (marked with star, top to bottom) first 20 synapomorphies shown, all other can be found on Appendix Table B.1.

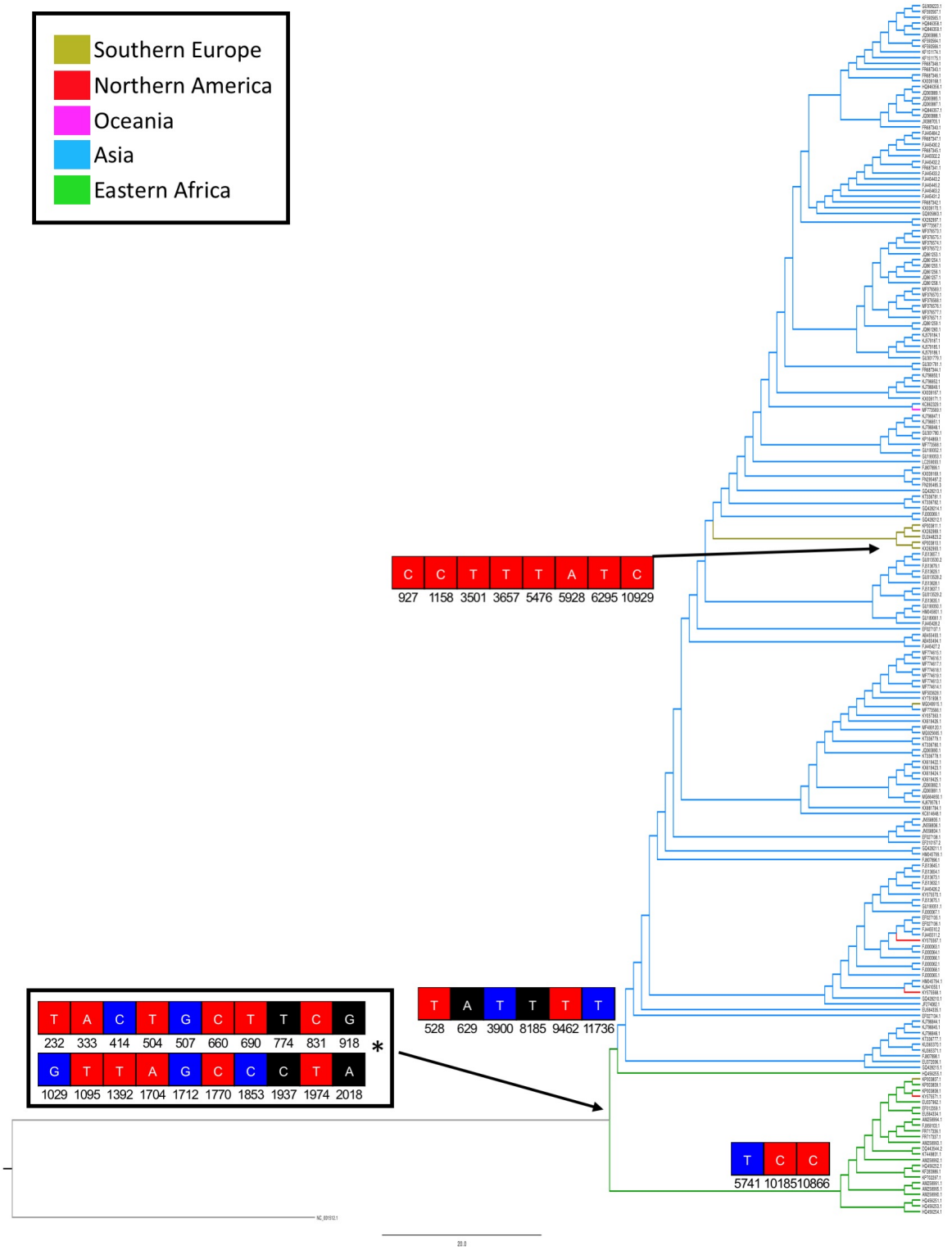

Figure 2.6: Chikungunya virus subtree - Eastern Africa / Indian Ocean Lineage clade - unique nucleic synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, Blue cells are non-unique, homoplastic synapomorphies. Node 719 (marked with star) first 20 synapomorphies shown, all other can be found on Appendix Table B.1.

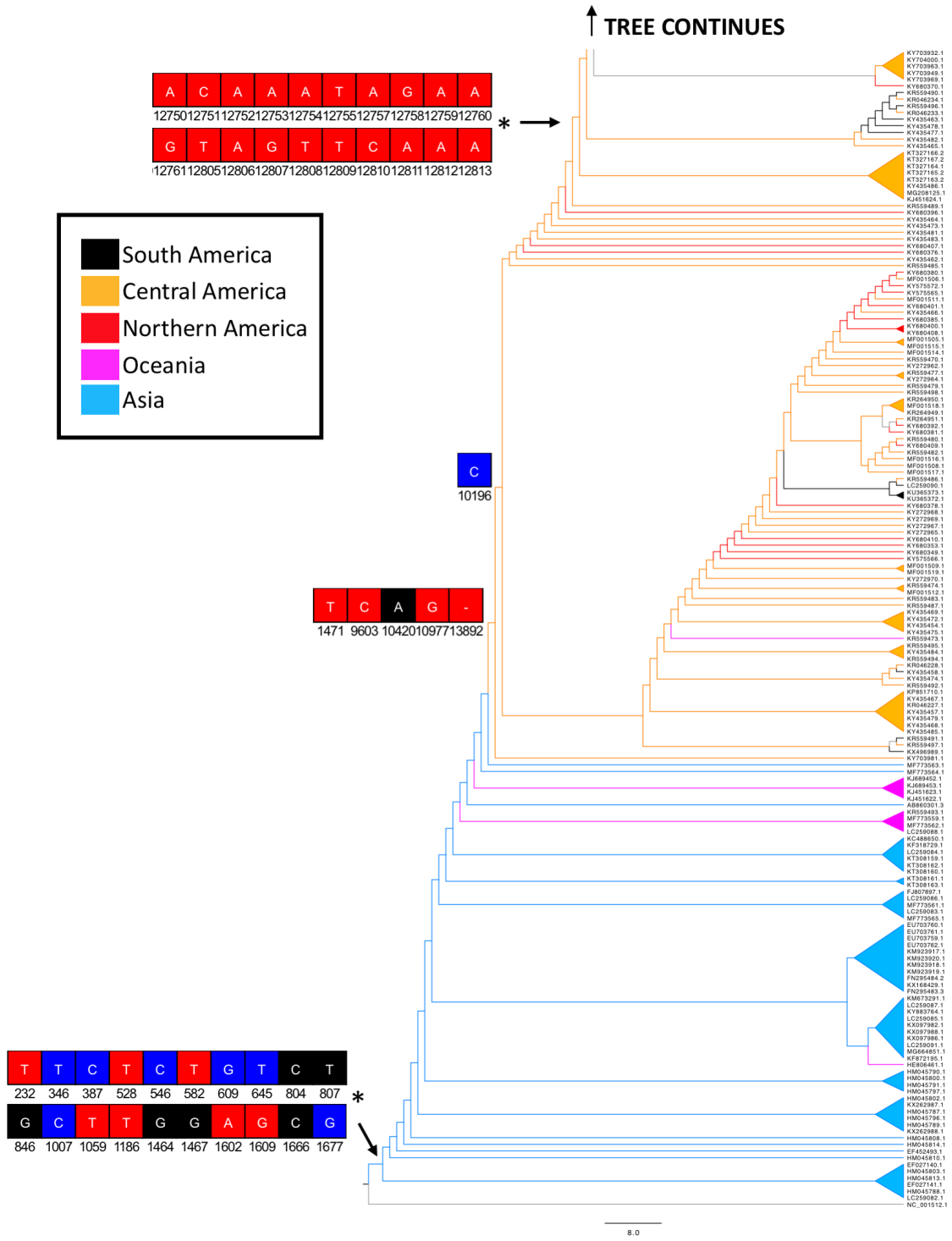

Figure 2.7: Chikungunya virus subtree - Asian Urban Lineage - unique nucleic synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, Blue cells are non-unique, homoplastic synapomorphies. Nodes 955 and 745 (marked with star, top to bottom) first 20 synapomorphies shown, all other can be found on Appendix Table B.1.

2.3.2.2 Zika virus

For ZIKV, the majority of strains grouped on the phylogenetic tree in two major current clades and an old African group (Figure 2.3. The oldest African strains were isolated from multiple *Aedes* and monkey hosts and do not form a monophyletic clade. All modern strains were isolates from humans. Asian strains have historically moved from Africa and group in two sister clades, one Asian-Pacific-American clade which includes the strains that caused the 2015-2017 ZIKV epidemic in the Americas and one recent clade in Asia. Like I have done for CHIKV, I split these major clades into subtrees in order to better understand the relationship between the taxa and map important synapomorphies:

1. African strains and Asian clade - Oldest African strains were grouped with Asian sequences that form a new monophyletic clade. Oldest African sequences were kept on this subtree only for easier visualization, as they are ancestral to both new Asian and American clades (Figure 2.8).

2. Asia Pacific American clade - Sister clade to the Asian clade, encompass island hopping sequences from Oceania and Southeastern Asia and the sequences from the 2015-2017 ZIKV outbreak in the Americas (Figure 2.13).

For ZIKV, all strains had annotated polyprotein, thus amino acid sequences were used to map synapomorphies given that looking on changes on amino acid sequence help to visualize the possible impact on protein function and/or structure. Only synapomorphies on key nodes (new clade/lineage based on geography) were evaluated.

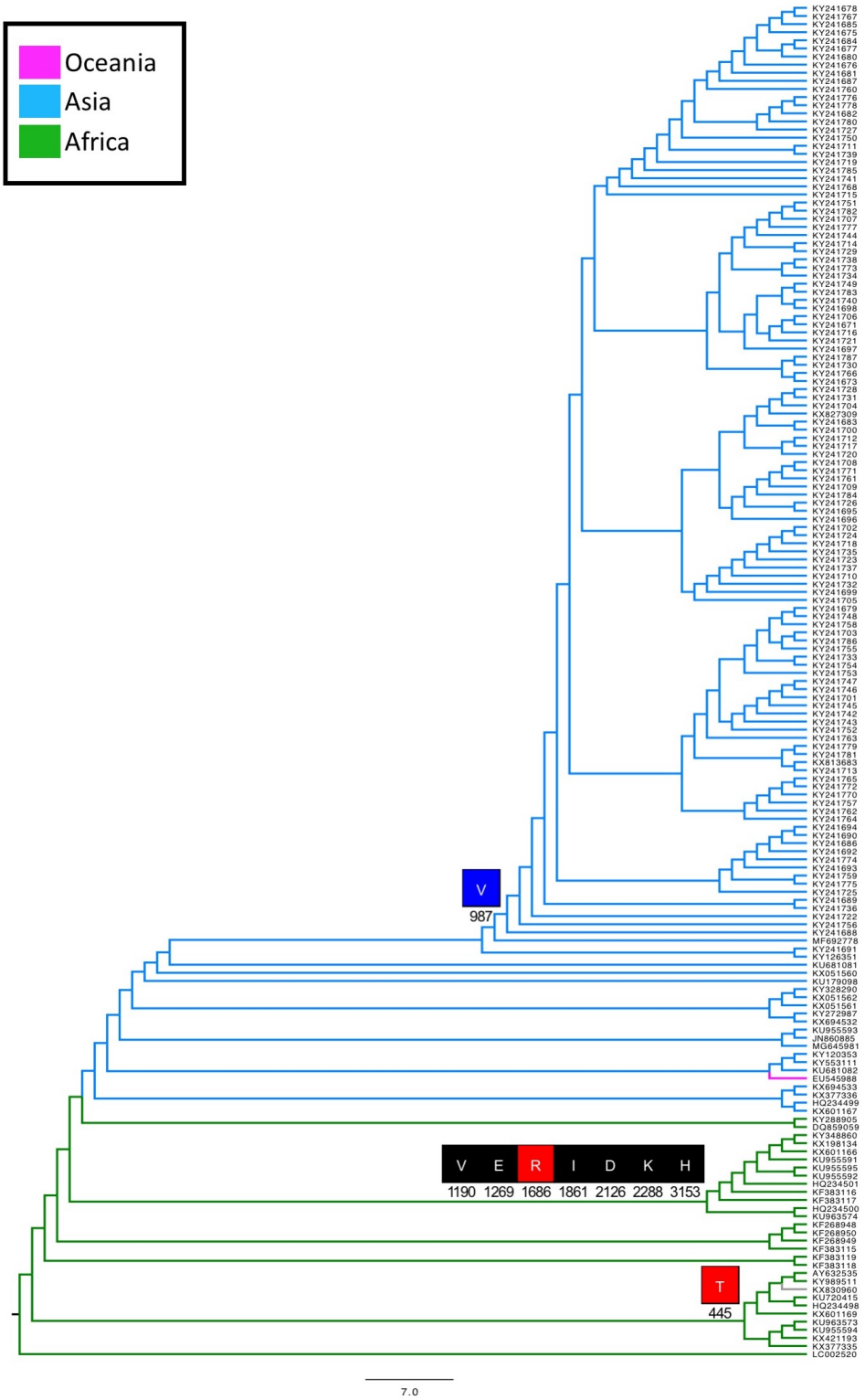

Figure 2.8: Zika virus subtree - African strains and Asian lineage - with mapped amino acid synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, Blue cells are non-unique, homoplastic synapomorphies.

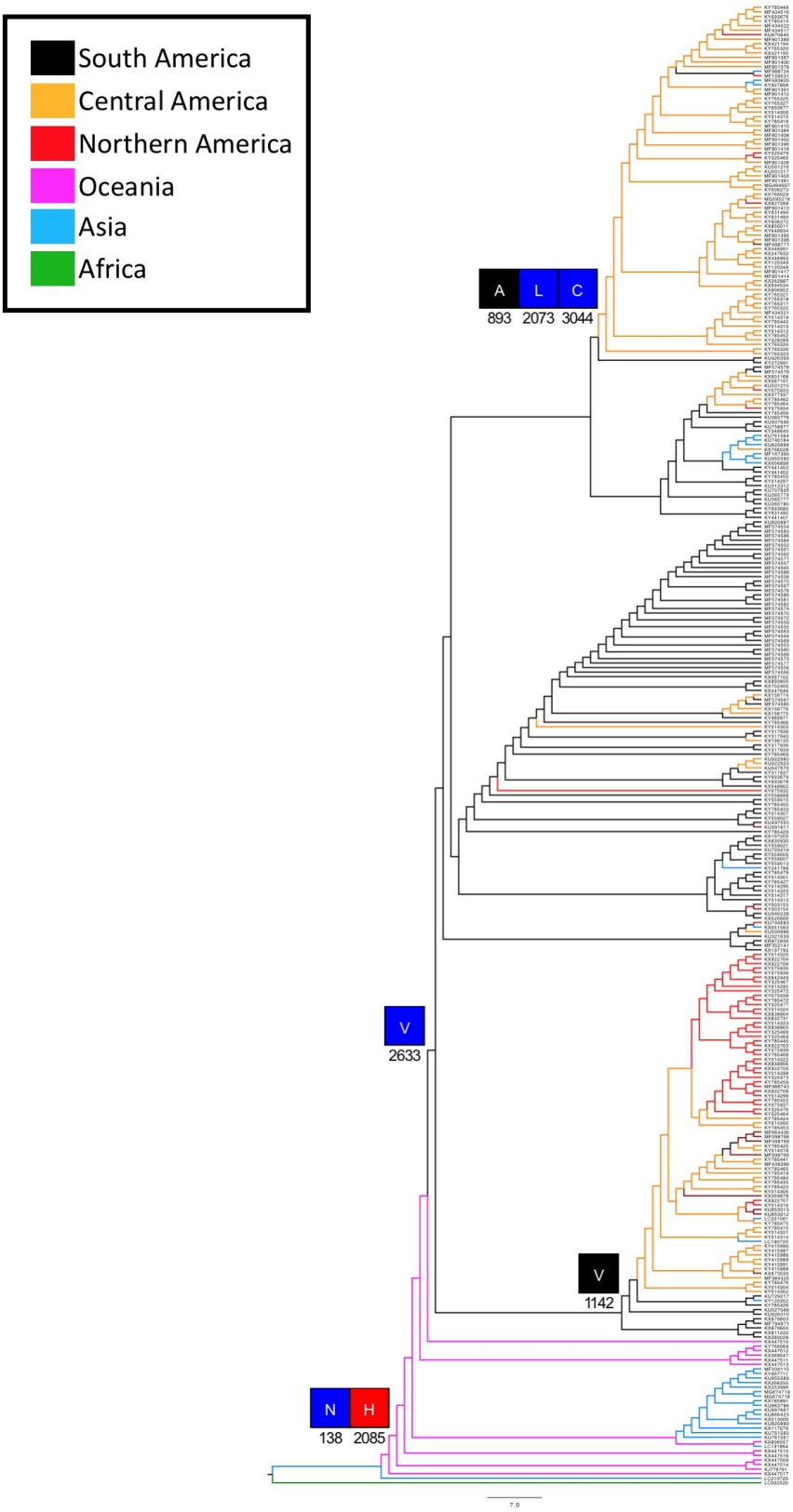

Figure 2.9: Zika virus subtree - Asia Pacific American lineage - with mapped amino acid synapomorphies. Black cells are unique, non-homoplastic synapomorphies, Red cells are unique, homoplastic synapomorphies, Blue cells are non-unique, homoplastic synapomorphies.

2.3.3 The global spread of Chikungunya and Zika viruses

The spread of this two distinct viruses mainly transmitted by the same vector, *Aedes aegypti* bring the question of whereas they follow a common trajectory in the world on specific lineages. For the purpose of visualizing their trajectories I split both viruses in two major and current clades. I split CHIKV viral strains in a clade with the Asian Urban Lineage, which spread from Asia and Oceania to Central and Northern America (Figure 2.10), and a clade that includes EA-IOL, MA-SA and Southern African lineages (Figure 2.11). The CHIKV West African lineage was excluded from this analysis given that its geographic spread is restricted to Western Africa. I split ZIKV strains in the Asia Pacific American lineage (Figure 2.13) and the clade that includes African strains and the Asian clade (Figure 2.12).

By georeferencing all the individual strains and also including the collection date and merging that information with the phylogenetic trees I was able to recreate the spread of the strains. This was done based on the full genome data and plotted into the globe utilizing the interactive tool Nvector.

2.3.3.1 Chikungunya virus

The spread of CHIKV is mainly due to the expansion of two major clades, although one of the major clades has expanded in a single direction (Asian Urban Lineage) and the other have multiple directionality (Middle Africa and Eastern Africa moving to different locations). I was able to rebuild the spread for CHIKV for both clades and took specific snapshots that encompass important movements on the spread of the virus.

For CHIKV Asian Urban lineage, the regions and time periods are the following:

1966 - Movement from India to Southern Asia;

2002 - Movement of the strain in Southern Asia;

2005 - Strain reaches America for the first time;

2010 - More spread in Asia;

2012 - Strain keeps spreading in Asia and Oceania;

2013 - Jump from Southeast Asia to Americas;

2014 - First boom in the Americas;

2015 - Spread from Americas to Pacific Islands;

2017 - Massive spread in the Americas.

For CHIKV African Asian strains the regions and time periods are the following:

1985 - Sequences only in Africa;

1986 - First jump out of Africa to Asia (India);

2005 - Movement in Africa, in 1995 first sequence that shows up in the US, traveler case [70];

2007 - Chikungunya sequence from France [71] and Chikungunya in Italy [72];

2010 - This Chikungunya strain keeps spreading more into southern Asia;

2012 - More spread in Southern Asia;

2014 - This Chikungunya strain jumps from Africa to South America;

2015 - More spread in Southern Asia, not much anywhere else;

2016 - Strain is found in the Caribbean countries;

2017 - 2 pictures, shows the overall spread of the African/Asian strain given the current dataset, strain present in Oceania.

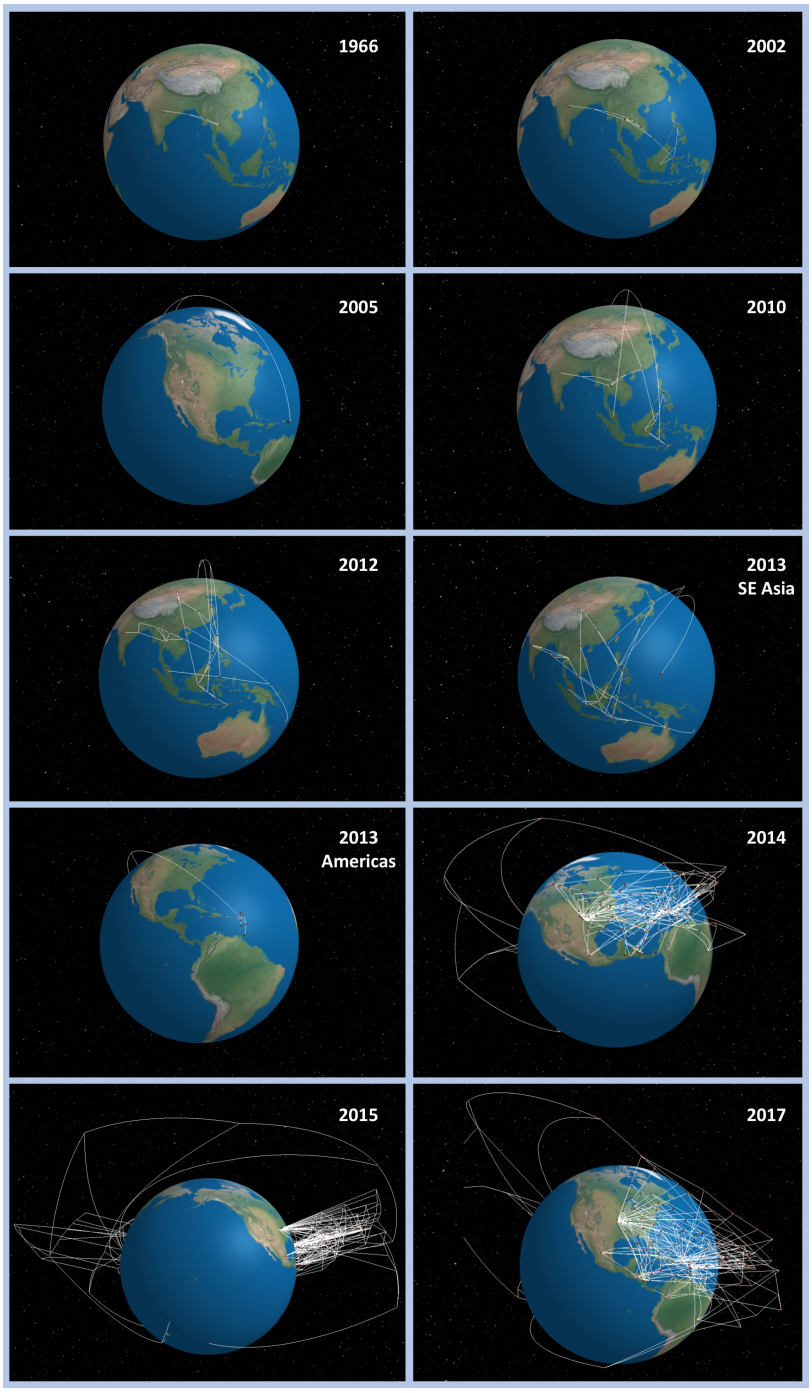

Figure 2.10: Rendering of the spread of the Chikungunya virus Asian Urban lineage strain across the globe from 1966 to 2017 based on the 410 taxa subtree. Each panel represents the geographical movement of the virus in a specific region during given time period. Red dots represent where the samples of the virus where collected and white lines show the geographical movement of the virus across the globe.

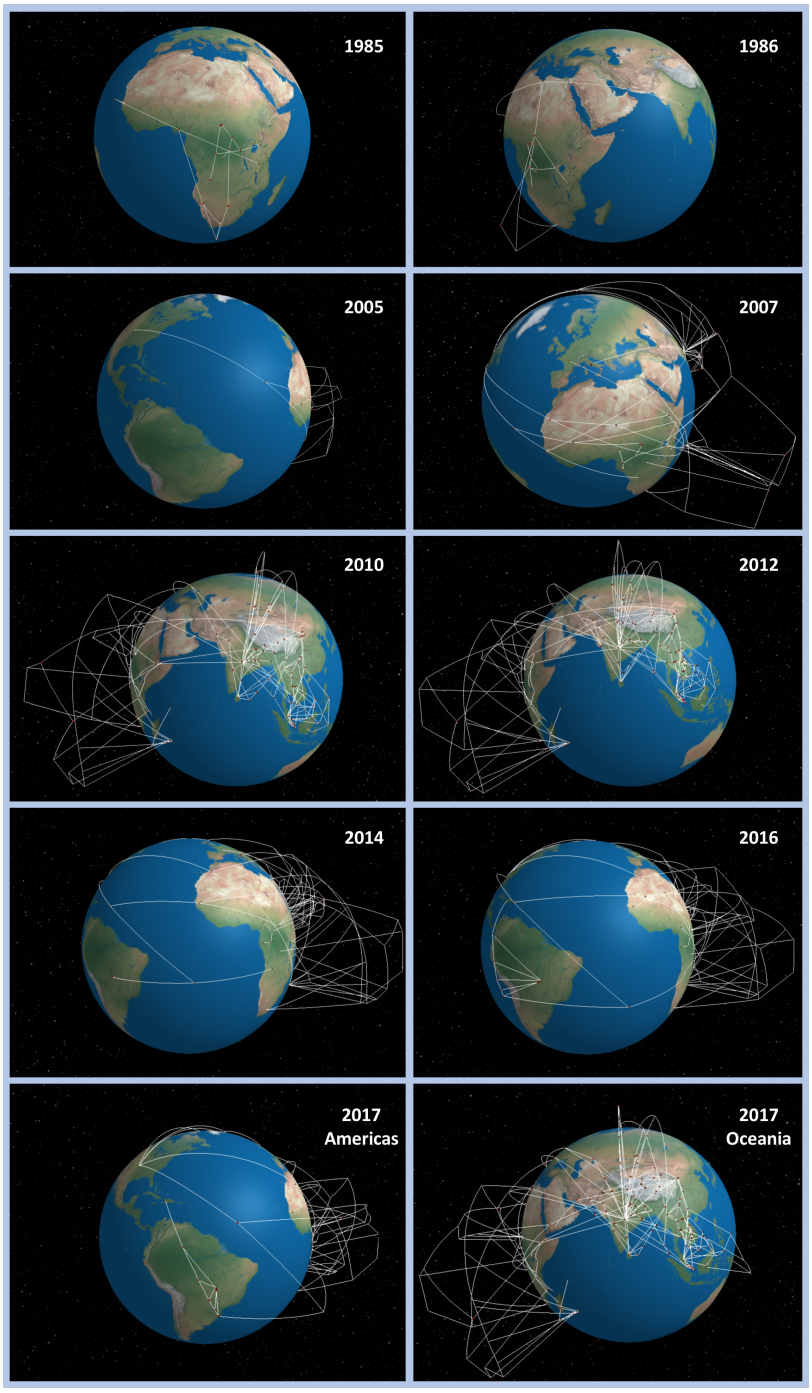

Figure 2.11: Rendering of the spread of the Chikungunya virus from Eastern and Central Africa across the globe from 1985 to 2017 based on the 265 taxa subtree. Each panel represents the geographical movement of the virus in a specific region during given time period. Red dots represent where the samples of the virus where collected and white lines show the geographical movement of the virus across the globe.

2.3.3.2 Zika virus

The ZIKV spread is intriguing as it was first isolated in the late 40's in Africa and mid 60's in Asia, but have a giant gap before being seen again for both lineages. The phylogenetic tree subsets rendered on the globe can provide a good perspective on how this gap due to lack of good surveillance, diagnostics technologies and ultimately sampling prevented a better outbreak response.

For ZIKV Asian American clade, the regions and time periods are the following:

1947 - ZIKV first sequence isolated in Africa;

1966 - First sequence isolated in Asia (serological evidence suggests earlier);

— TIME GAP —

2012 - Outbreak in Micronesia;

2013 - Outbreak in French Polynesia (first cases of Microcephaly recorded);

2015 - Strain reaches South America and rapidly spread to Central America;

2017 - Strain is found in North, Central and South America.

For the more recent ZIKV Asian clade, starting with the Asian strain from 1966, the regions and time periods are the following:

1966 - ZIKV Asian strain had its first sequence isolated in Asia (serological evidence suggests earlier);

— TIME GAP —

2011 - Spread of ZIKV in Southern Asia and Pacific Islands;

2014 - More movement of virus in Southern Asia;

2017 - Strain differentiates itself from others forming the Asian strain.

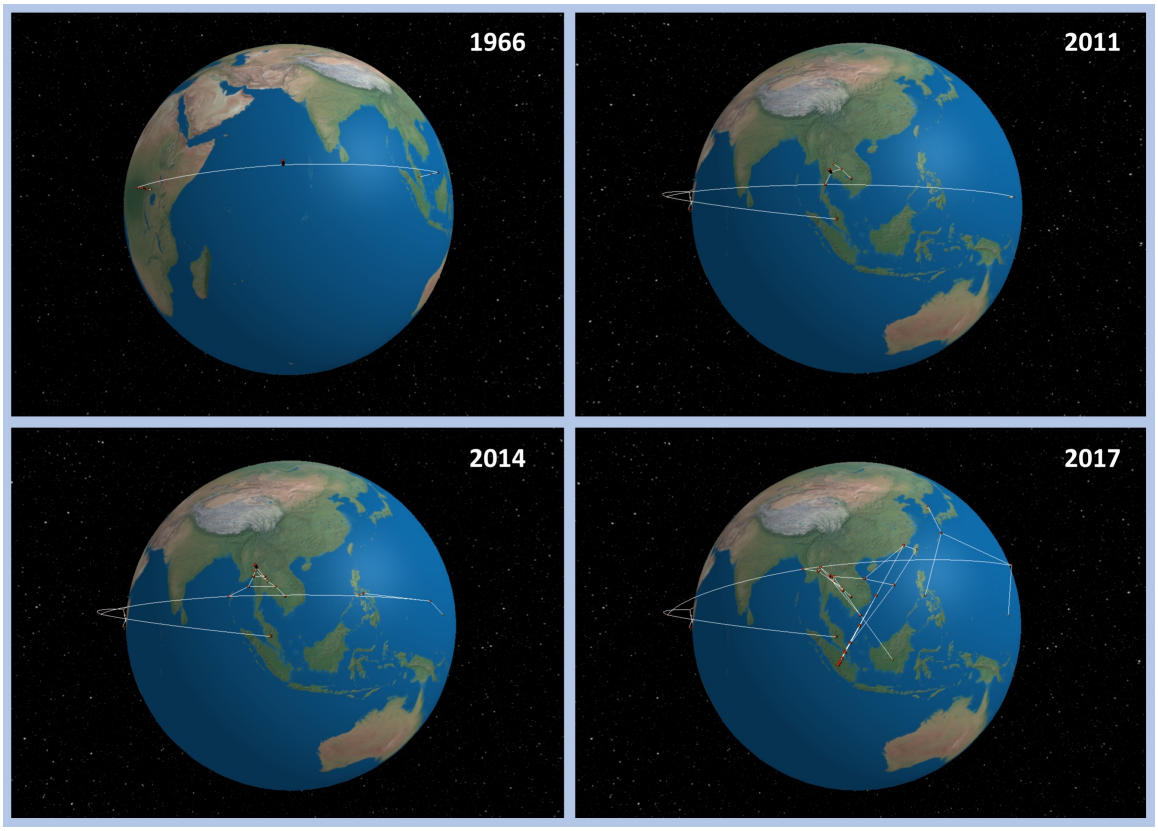

Figure 2.12: Rendering of the spread of the Zika virus Asian lineage across Eastern Hemisphere from 1966 to 2017 based on the 167 taxa subtree. Each panel represents the geographical movement of the virus in a specific region during given time period. Red dots represent where the samples of the virus where collected and white lines show the geographical movement of the virus across the globe.

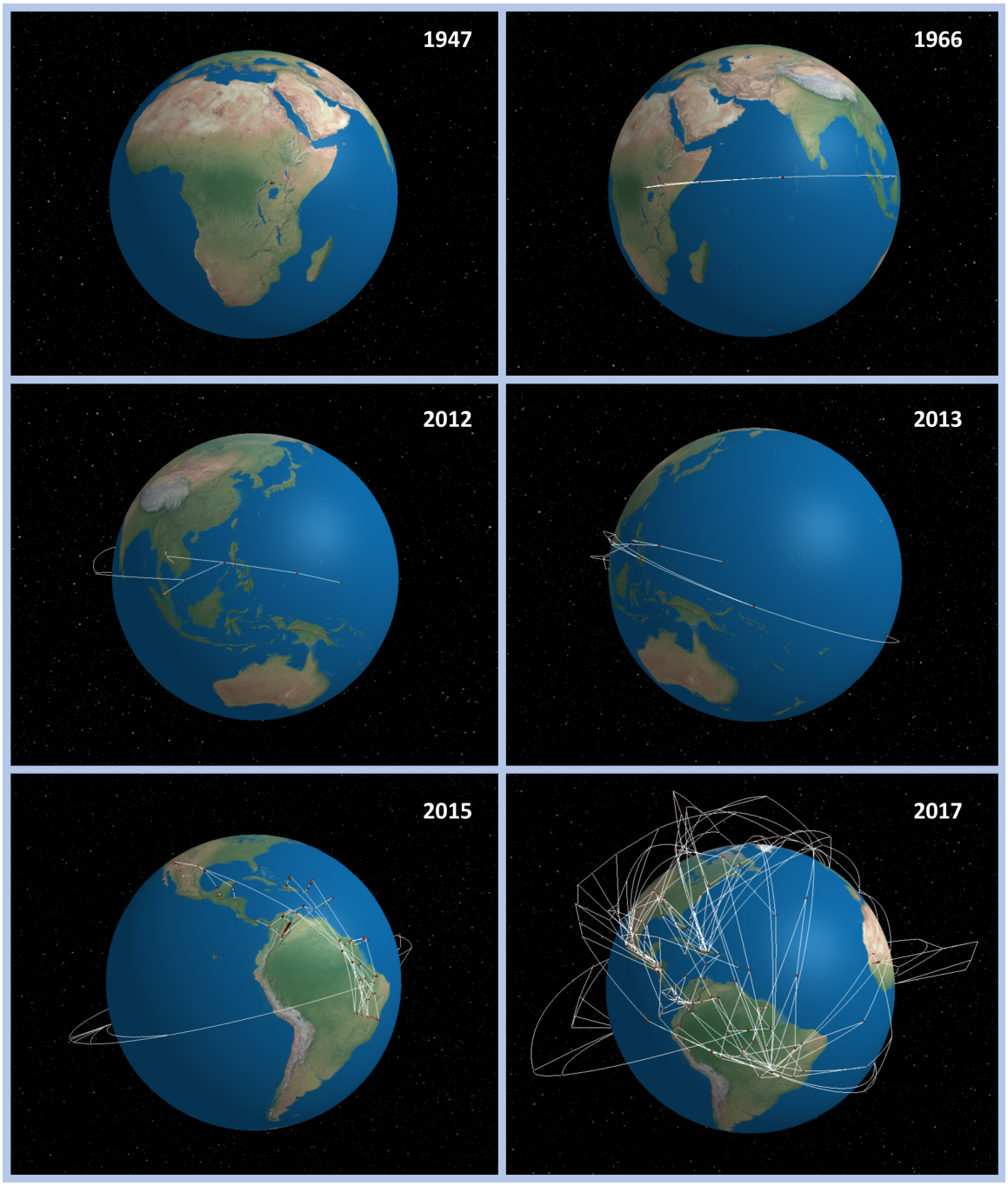

Figure 2.13: Rendering of the spread of the Zika virus Asia Pacific American lineage across the globe from 1947 to 2017 based on the 374 taxa subtree. Each panel represents the geographical movement of the virus in a specific region during given time period. Red dots represent where the samples of the virus where collected and white lines show the geographical movement of the virus across the globe.

2.3.4 Untranslated Regions

UTRScan and RegRNA 2.0 searches yielded a total of 13012 hits on all ZIKV and CHIKV 3 and 5'UTRs 2.1. All available UTR sequences for CHIKV and ZIKV were utilized on UTRScan as it consists of a matching pattern search algorithm that provides quick results. Prior to the search of elements of the UTRs on RegRNA 2.0 all redundant sequences were excluded as RegRNA 2.0 provides a more in-depth search within the UTR sequences.

Table 2.1: Summary of findings on CHIKV and ZIKV UTRs.

N	**Virus & Sequence**	**Database**	**N Input/ Hit Seqs**	**N of Hits**	**Description**
1	ZIKV 5'UTR	UTRdb	488 / 6	8	Terminal Oligopyrimidine Tract (TOP); Internal Ribosome Entry Site (IRES); Upstream Open Reading Frame (uORF)
2	ZIKV 3'UTR	UTRdb	487 / 449	736	Upstream Open Reading Frame (uORF); Internal Ribosome Entry Site (IRES)
3	CHIKV 5'UTR	UTRdb	281 / 0	0	N/A
4	CHIKV 3'UTR	UTRdb	281 / 226	805	Internal Ribosome Entry Site (IRES); Upstream Open Reading Frame (uORF); Polyadenylation Signal (PAS)
5	ZIKV 5'UTR	RegRNA 2.0	123* / 119	303	ighg2; cgamma2 - intron 1; ctnt, exon 5; Polyadenylation sites; MYB; Cdx-1; AP-1; Cart-1; Fra-1; NF-E2; c-Maf; MAFB; AP1; OC-2; GABP-alpha

Table 2.1 continued from previous page

N	Virus & Sequence	Database	N Input/ Hit Seqs	N of Hits	Description
5 cont'd	ZIKV 5'UTR	RegRNA 2.0	123* / 119	303	Musashi binding element (MBE); Hammerhead form III; CRE type 3 poliovirus; TAR HIV-1; u1107 KLHL1-antisense-RNA
6	ZIKV 3'UTR	RegRNA 2.0	271* / 271	6638	gh1: growth hormone 1, exon3; gh-1, exon 3; sc35 - exonic splicing enhancer; cftr, exon 12; brca1, exon18; ighg2 cgamma2; ctnt, exon 5; gh-1 intron 3; tau, exon 10; gh1:growth hormone 1, intron3; Polyadenylation sites; Ik-2; AP-1; c-Myc:Max
6 cont'd	ZIKV 3'UTR	RegRNA 2.0	271* / 271	6638	Pax-4; MAF; HOXA7; Kid3; NFAT1; Neuro D; NF-AT4; EHF; Elf5; TEL1; Fra-1; myogenin; NR1B2; Elk-1
6 cont'd	ZIKV 3'UTR	RegRNA 2.0	271* / 271	6638	ETF; AP-2; CP2/LBP-1c/LSF; NF-AT2; ING4; ETV7; C/EBPgamma; BEN; CRX; GATA-1; LRF; LXRalpha:RXRalpha; AP-4; FAC1

Table 2.1 continued from previous page

N	Virus & Sequence	Database	N Input/ Hit Seqs	N of Hits	Description
6 cont'd	ZIKV 3'UTR	RegRNA 2.0	271* / 271	6638	Musashi binding element (MBE); Hammerhead form III; CRE type 3 poliovirus; TAR HIV-1; RF00525; RF00185; FR264255/Flavivirus DB element; FR156746/Flavivirus DB element; FR149872/Flavivirus DB element; FR063977/Flavivirus DB element; FR369244/Flavivirus DB element; FR284296/Flavivirus DB element; FR268518/Flavivirus DB element; FR133895/Flavivirus DB element
6 cont'd	ZIKV 3'UTR	RegRNA 2.0	271* / 271	6638	FR155608/Flavivirus DB element; hsa-miR-4494; hsa-miR-637

Table 2.1 continued from previous page

N	Virus & Sequence	Database	N Input/ Hit Seqs	N of Hits	Description
7	CHIKV 5'UTR	RegRNA 2.0	59* / 56	208	ighg2; cgamma2; Ncx; PARP; CDP; OC-2; GTF2IRD1-isoform2; Pax-8; FR293439/Togavirus 5' plus strand cis-regulatory element; FR375184/Togavirus 5' plus strand cis-regulatory element; FR005458/Togavirus 5' plus strand cis-regulatory element

In 2016, I evaluated the UTRs of ZIKV and identified utilizing UTRScan a mutation prior to an element identified on the 3'UTR called Musashi Binding Element (MBE) of ZIKV that was conserved on what was called at the time the Asia Pacific American lineage [32] (Figure 2.14). Calculating the binding energy according to Zearfoss et al., 2014 [1] I was able to see an increase on the binding affinity of that motif on human Musashi proteins (Table 2.2) [2].

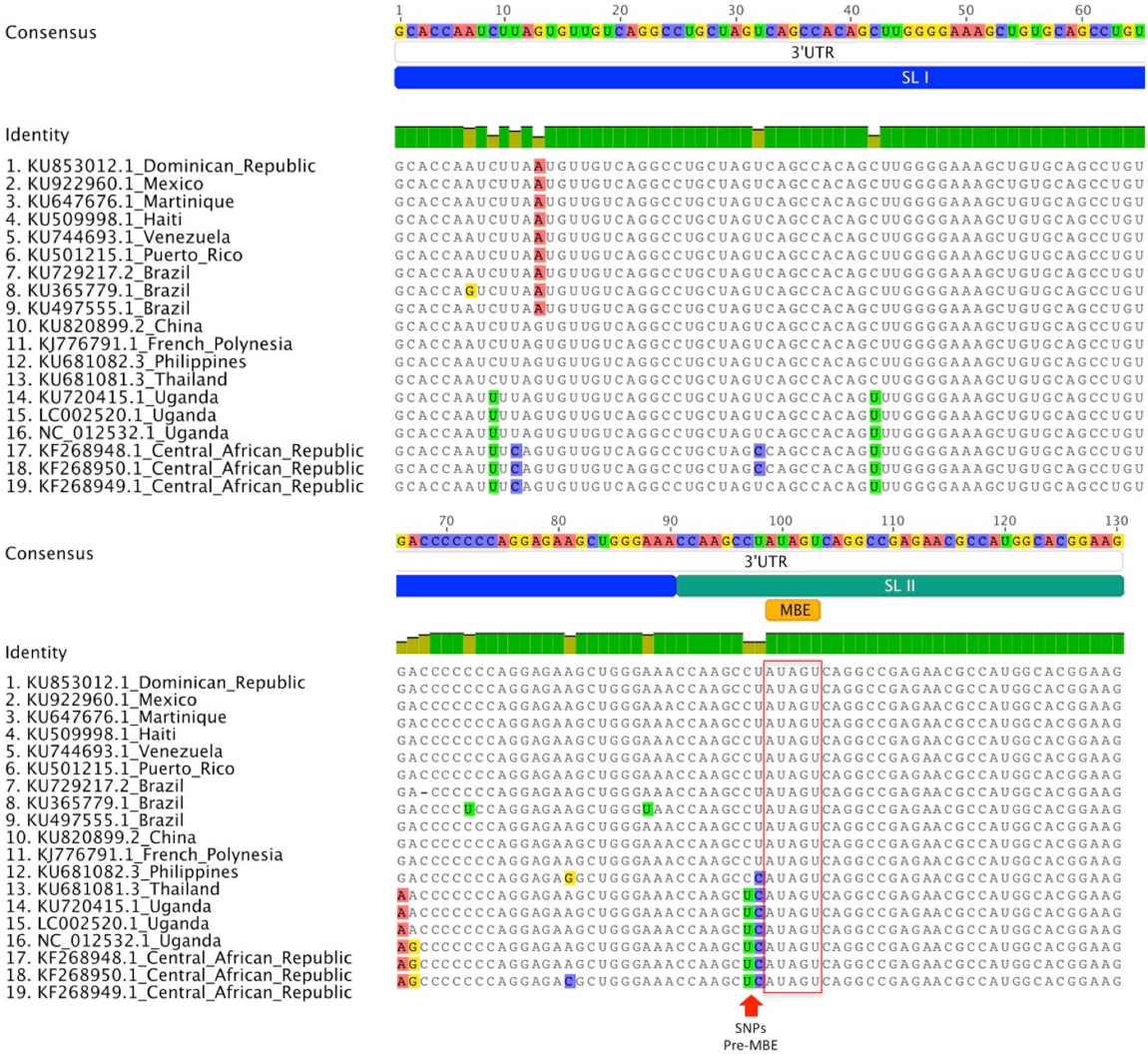

Figure 2.14: Alignment of 3'UTR of ZIKV with associated mutations over geographic spread [32].

Table 2.2: Summary of predicted binding free energy change ($\Delta\Delta$ G kcal/mol) attributable to corresponding ZIKV MBE sequence relative to comparison sequence per Zearfoss et al., 2014 [1, 2].

Predicted binding free energy change	**NC012532 Uganda**	**LC002520 Uganda**	**KU681081 Thailand**	**KU681082 Philippines**	**KU527068 Brazil**	**KU820899 China**
Mouse MSI1 RRMl-2 to Zika MBE	(-0.04)	(-0.04)	0	(-0.67)	(-0.67)	(-0.67)
Mouse MSI1 RRMl to Zika MBE	(-0.23)	(-0.23)	0	(-0.22)	(-0.22)	(-0.22)
Human MSI2 RRMl-2 to Zika MBE	-0.03	-0.03	0	(-0.64)	(-0.64)	(-0.64)
Drosophila MSI RRM1-2 to Zika MBE	(-0.04)	(-0.04)	0	(-0.15)	(-0.15)	(-0.15)
Drosophila MSI RRM1 to Zika MBE	(-0.04)	(-0.04)	0	(-0.09)	(-0.09)	(-0.09)

The MBE motif was not found in the latest search on UTRScan as the updated version of the tool has reduced the number of motifs it searches for, but the search utilizing RegRNA 2.0 identified the presence of MBE in both CHIKV and ZIKV 3'UTR as well as on ZIKV 5'UTR. For the purpose of identifying if MBE could also have a role in CHIKV pathogenesis, I calculated the opening energy of the MBE present on the 3'UTR of three CHIKV (NC_004162.2 African S27, KX262997 - Indian Ocean Lineage, KY680376.1 - American) and three ZIKV (NC_012532.1 - African, NC_035889.1 - American and KY24176.1 - Asian), each representing one of three the major clades for each virus (Figure 2.15). The more negative z scores indicate the accessibility of the trinucleotides. As expected according to previous studies, the Brazilian ZIKV isolate had the lowest z score, followed by Asian CHIKV (KX262997.1) and ZIKV (KY241761.1) isolates. The original ZIKV sequence, MR766 from Uganda (NC_012532.1) had similar z-scores to the American (KY680376.1) and the African (NC_004162.2) CHIKV isolates. A summary of these and all results of this chapter can be found on Table 2.3.

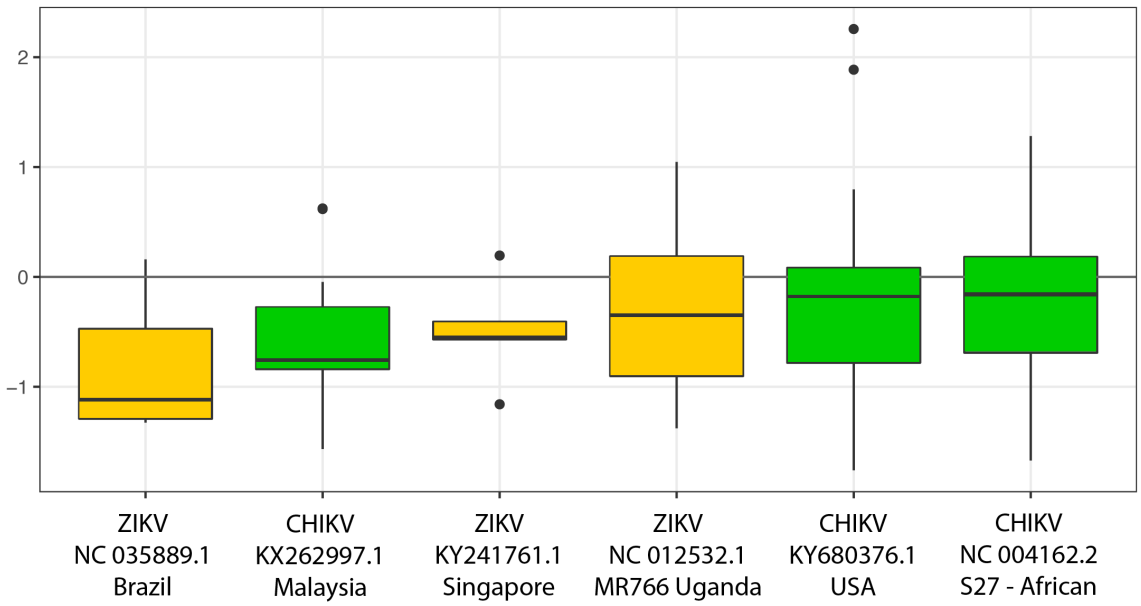

Figure 2.15: Z scores of MBEs the opening energies of MBEs on 3'UTR sequences of the major clades of Chikungunya and Zika viruses.

Table 2.3: Summary of chapter 2 results.

	Chikungunya virus	**Zika virus**
Phylogenetics	3 major clades / 4 independent evolving lineages	2 major clades (Singapore sequencing is responsible for Asian clade)
Nvector	Origin: Africa Spread: Asia to America / Africa to America / Africa to Asia	Origin: Africa Spread: Africa >Asia >Island hopping to America
Synapomorphies	Major insertion on UTR of Asian Urban lineage	Minor conserved mutations across genome
UTR analyses	Musashi Binding Element is mostly double stranded, not available	Musashi Binding Element is the reason for presence of Zika in spermatogenesis and neural cells.

2.4 Discussion

2.4.1 Phylogeny and spread of Chikungunya and Zika viruses

The analyses of ZIKV utilizing 491 isolates confirm that the ZIKV lineage found in the Americas descended from Southeastern Asia, island hopping until it reached Brazil from French Polynesia and from there it spread to other countries in South, Central and North America. Although not largely spread in Asia, a second major clade can be visualized out of Asia, which can be inferred as a new modern Asian lineage. Further evidence has to be presented to confirm if these are isolated events or it is really a new lineage that is currently circulating in Asia given that the vast majority of sequences in the Asian lineage were sequenced during a single outbreak in Singapore[73]. The spread of ZIKV seen on this work agree with other recent studies that propose a silent spread of ZIKV for years in Asia prior to the virus reaching the Americas [74].

By looking at the branch lengths of the ZIKV phylogenetic tree it can be seen that ZIKV African sequences are considerably distant from the Asian lineages (Appendix

Figure A.2). The lack of epidemiological history and genomic data of ZIKV in both Africa and Asia for over 50 years remains a puzzle that may not ever be solved. The lack of a severe disease phenotype and serological cross-reactivity on antibody assays utilized in the past to classify patients *Flavivirus* infections let, with exception to outbreaks in French Polynesia, to ZIKV went unnoticed until microcephaly was observed in Brazil in 2015 [75, 76].

The first cases of ZIKV return in Africa were observed in October 2015 in Cape Verde. Although no sequence data has become publicly available, the World Health Organization (WHO) confirmed that these cases were imported from the Americas. Moreover, this was the first time microcephaly was associated with ZIKV in Africa [77]. The presence of this outbreak in Africa for the first time in over 50 years makes ZIKV have completed for the first time the circumnavigation of the Earth [2].

While WNV transmission has become more common in Europe, weather and vector distribution have been apparently protecting Europe from having ZIKV endemic in the region. Different from WNV which is transmitted by the *Culex* mosquito, ZIKV main vector is *Aedes aegypti*, not highly established in Europe as the main mosquito of this genus present in Europe is *Aedes albopictus*, which is known to have a competition advantage over *Aedes aegypti* [78]. This may explain why although there were over 2000 cases reported in 21 European Union countries between June 2015 and January 2017, almost all cases were travel cases where individuals got infected while traveling outside Europe, and a few cases were sexually transmitted [79].

Similarly to ZIKV, CHIKV presents two major clades, although it possesses a complete different spread pattern. In the case of ZIKV, the spread to America seems clear now, but the question on how it spread in Asia still open. In CHIKV, better identification of the disease phenotype have kept these gaps of information smaller. CHIKV had multiple waves of infection over the past two decades, coming out of Eastern and Middle Africa, causing outbreaks in Southeastern Asia (Thailand, Malaysia, etc),

Southern Asia (India) and the Americas. The clades observed in the analysis match those found in the literature, although it is interesting to see that with the amount of new sequences what was once called ECSA, East Central South African lineage, is not really a single lineage, at least in the sense of a synonym to a clade, or to describe the series of descendants of an ancestor up to a specific terminal without side branchings. For example, Middle Africa strains gave origin to the outbreak in South America, while Eastern African strains gave origin to the Asian Lineages. Also, two are the lineages in Asia, one called Indian Ocean Lineage, which, up to the moment remains exclusive of the Asian continent, with exception of a couple outbreaks it has caused in Europe, and a second lineage called Asian Urban which reached the American continent, possibly going to establish overtime a new American Lineage. Much still has to be discovered as more isolates are sequenced, helping to better understand the dynamics of outbreaks. CHIKV has also caused recent outbreaks in Europe caused by the IOL, and different from ZIKV, CHIKV has been found to be transmitted effectively in humans by *Aedes albopictus* [80].

It is important to raise the significance of a definition of lineage when discussing the evolution of a virus. Much has been discussed as how species should be delineated. In 1977, Wiley proposed that the definition of species should be that "A species is a single lineage of ancestral descendant populations of organisms which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate" [81]. By lineages, Wiley utilized the concept of an ancestral-descendant sequence of populations [82]. At the same time, he implied that species are "historical, temporal, and spatial entities". When looking at viruses, Peterson criticizes current system of taxonomy of viruses, and mentions that one criterion alone should dominate viral classification, the evolutionary independence of evolving lineages [83]. Looking at viral lineages, and treating each strain as an individual to be classified, keeping the idea of independence in mind, currently it seems that CHIKV

and ZIKV have not had enough time to diverge from their ancestral sequences in Asia to be considered alone American lineages. Independent of the American lineage, Asian has a modern lineage that has circulated in Southeastern Asia. Thus, with the establishment of endemic regions for viruses in geographic distinct and isolated regions, isolates will have to be assigned to their correct and evolutionary closest endemic region. Peterson criticizes the use of endemic regions to classify species, although this could be a good pointer in association with evolutionary distance to classify lineages.

2.4.2 Genomic epidemiology of Chikungunya and Zika viruses

2.4.3 Chikungunya virus

The Western African clade had the largest number of synapomorphies compared to the other major clades. It is expected that this clade diverge from its sister clade over time as this is the only clade which is geographically isolated from other regions with no transmission events to or from other geographic regions being observed.

All other lineages observed although forming separate clades coexist in Southeastern Asia and may have started to coexist in the Americas, which may allow opportunities for genetic recombination between lineages [84]. An insertion of 30 nucleotides on a node on the Central North American outbreaks also points to the importance of the UTRs on the evolution of CHIKV. As previous papers mentioned before, the length and structure of CHIKV UTRs have a strong role in the evolution of the virus [85]. The reason why IOL has been effectively being transmitted in Europe has to do with a single Alanine to Valine mutation on the Envelope protein (E1-A226V) which affects vector specificity (better competency on *Aedes albopictus* and consequently its endemic potential [24].

Unfortunately missing vector information due to isolation in human patients as well as no clinical data associated with the sequence do not permit me to investigate further the effect of the observed mutations and serve here as the starting point

to compare lineages and investigate why they vary in pathology and have different patterns of spread [17]. Thus, further studies have to focus on acquiring accurate metadata from the updated datasets to identify if there are novel trends associated with vector specificity and pathogenicity.

2.4.3.1 Zika virus

The ZIKV sequences from both Asian and Asia-Pacific-American clades are highly conserved, novel mutations been observed in the literature but there is too much debate on whether they mean an actual change on the behavior of the disease or not [86, 87]. The Asian clade has only one non-unique homoplastic synapomorphy mapped. Given that it is just seen as a major clade due to 116 sequences from a single outbreak in Singapore, there is a need of time to see if it will actually evolve into a major clade. Sequences from multiple hosts also add a confounding effect when comparing old African isolate sequences with modern sequences.

Much discussion is also brought on whether microcephaly and other phenotypes are actual novel and belong to the current strains or this been just a phenotype overlooked in the past due to lack of surveillance and the co-circulation of other diseases with severe phenotypes. The observed low number of conserved synapomorphies shared within the major ZIKV clades show that there was no major accumulation of mutations within the polyprotein that drove change in disease behavior. Rather than changes in the polyprotein, this study point to changes in untranslated regions as a possible explanation to ZIKV change in behavior.

2.4.4 Untranslated Regions and their role in pathogenesis

Many researchers overlooked over the translated regions of the genome remain the untranslated regions and their structures. Mainly for lack of knowledge or information on how to proceed with the analysis, researchers seem to focus on changes on conserved sequences rather than changes in not-so-conserved sequences that have

conserved structure, such as UTRs [88].

The 3'UTRs in arboviruses determine replication and virus host range [89, 90]. Multiple motifs were mapped to CHIKV and ZIKV UTRs utilizing UTRscan and RegRNA 2.0. Over all motifs found, MBE was selected due to its binding target, the protein Musashi, which is known to be highly present in spermatogenesis and neurological development among other tissues [91, 92]. This binding target presence in these specific tissues correlate with recent observed pathology and flagged MBE as a potential mechanism by which ZIKV changed tropism and may be causing birth defects.

Differences within ZIKV lineages been observed, behavior of ZIKV UTR region is different from genomic and ZIKV has the lowest opening energy for MBE among all flaviviruses [93]. When performing a comparison between CHIKV and ZIKV, I found that most ZIKV sequences have the lowest opening energy. In 2018, Platt et al. [94] investigated whether two flaviviruses, WNV, Powassan virus (POWV), and two alphaviruses, CHIKV and Mayaro virus (MAYV) could cause fetal demise in mice and found that, although all viruses could cause placental infection, only WNV and POWV cause fetal demise. Nonetheless, CHIKV has the ability to cause placental infection and can generate complications of neonatal illness [95].

The presence or absence of MBEs does not exclude the possibility of congenital defects on other species, it only serves as a pointer to whether this could be or not the mechanism by which the defects are being caused [93]. Given my results and that previous literature points to severe congenital defects been mainly observed within flaviviruses, one can assume for the moment that the MBE role on pathogenicity is exclusive to flaviviruses, if not only ZIKV. Further studies have to be performed on alphaviruses to rule what is the MBE role on these sequences given the presence. Nevertheless, my initial results point to MBEs not being structurally available for binding in CHIKV. When looking in depth on ZIKV, there is a mutation prior to

MBE that seems to increase the binding affinity of MBE to Human Musashi [32, 2]. Chavali et al [65] have shown experimentally that Musashi-1 protein is highly expressed in neural precursors and could explain the vulnerability of those cells to the virus, endorsing the bioinformatics findings done here.

CHAPTER 3: TRANSMISSION NETWORKS

3.1 Introduction

Recent outbreaks of viruses from the *Alphavirus* and *Flavivirus* families have raised the need of a better monitoring system in place to assist on the development of abatement strategies. As mentioned previously, ZIKV have emerged crossing from Asia to the Americas via islands of the Pacific Ocean in 2015, CHIKV have also recently reached the Americas and other flaviviruses such as YFV and DENV have been causing recent epidemics. The 2017 YFV outbreak in Brazil is ongoing although no major reports have been issued since March 2018 [96]. Dengue, with its four serotypes is a recurrent issue in tropical countries, more especially those with areas with poor sanitary infrastructure (Table 3.1).

Medical countermeasures and response to viral outbreaks have been shaped through the advance of new technologies. High throughput sequencing, combined to increased computer power, bring an avenue of new tools to analyze the evolution of viruses. The data generated nowadays, allow us to infer the relationship between samples of the same virus collected from individuals in different parts of the world and with that information in hands, recreate the steps it took for a virus to reach a certain location.

The study of where viruses originate and how they spread (e.g. among various hosts) can be made by combining the knowledge of phylogenetics with network theory. In this effort the information is extracted from phylogenetic relationships and metadata to build transmission network graphs. These graphs can demonstrate not only the source of outbreaks but also key geographic regions, or hosts, or food sources that facilitate the spread of the disease.

Increased viral surveillance, as well rapid sequencing of pathogens from infected

patients worldwide is the current path that will allow the more granular studies required to investigate emerging diseases as they evolve and spread. With genetic data available the pipeline described on the methods of this objective will allow us to better understand disease outbreaks, which can assist health authorities to respond more efficiently to outbreaks and to plan methods for containment of diseases.

3.1.1 Chikungunya virus transmission cycle and historical data

Summarizing what has been described in objective one, Chikungunya virus was first isolated in Tanzania in 1952 [5]. The virus originated in Africa and it is believed to have first spread from Africa to other parts of the world through sailing ships in the 18th Century [19]. The virus was first found in the Americas in October, 2013 in the island of Saint Martin [6]. In the literature, the virus comprises of at least four lineages: Asian lineage, IOL, ECSA and West African [17].

CHIKV transmission cycle is divided in sylvatic and urban, having on its sylvatic cycle non-human primates and mosquito vectors and the urban cycle the mosquito-human-mosquito cycle. Occasionally, spill overs from the sylvatic cycle tend to cause epidemics in the urban cycle. It is believed that in Africa, the sylvatic cycle is the predominant life cycle of the virus, whereas in Asia the urban cycle dominates [97]. The first emergence in urban cycle is estimated between 1879 and 1956 when ECSA lineage went to Asia.

As of December 2017, on the PAHO Epidemiological Week 51 for the Americas over 120 thousand cases of confirmed CHIKV cases were confirmed which include 101 deaths. Of those cases the country with the absolute majority of cases was in Brazil [98]. As of September 2017, over 100 cases were confirmed in the region of Lazio in Italy. As of August 2017, four cases were confirmed in France [99].

3.1.2 Dengue virus transmission cycle and historical data

DENV is believed to be a continuous burden in society stemming back several centuries. A first description of symptoms that resemble the disease can be found in a Chinese medical encyclopedia, which was published by the Chin Dynasty somewhere between 265-420 AD [100]. In the 1600's, epidemics with similar symptoms occurred in the West Indies and Central America [101]. In the 1700's, DENV reached the continental United States in North America and became a recurrent problem until its last recorded autochthonous outbreak in 1945 in New Orleans [102].

The DENV has four different serotypes, DENV-1, DENV-2, DENV-3 and DENV-4. The first two DENV (DENV-1 and DENV-2) were isolated in 1943 and 1945 from one infection in Japan and one in Hawaii, respectively [103]. The serotypes share approximately 65% similarity of their genome, similarity which can also be seen when comparing other flaviviruses such as WNV and Japanese encephalitis virus [104].

DENV is primarily transmitted by the *Aedes aegypti* mosquito and secondarily other *Aedes* species such as *Aedes albopictus*. DENV has a sylvatic cycle and an urban cycle. The sylvatic cycle consists of wild non-human primates hosts and various species of Aedes as vectors. The urban cycle consists of the mosquito-human-mosquito interaction [105]. Phylogenetic studies suggest that epidemic strains of DENV evolve periodically from sylvatic progenitor lineages [106, 107]. A recent study questions the current hypothesis given the limited amount of sampling at the time, Damodaran et al. [108] suggests that a reverse pattern occurs, sylvatic strains often emerge from epidemic strains, based on the same data as Wang et al., 2000 [106] and updated datasets.

Currently, approximately 50% of the world's population lives in areas at risk of DENV as defined by the presence of the vector mosquitoes. More than 125 countries are known to have endemic DENV. DENV is present in all regions of the world as classified by the WHO [109]. It is estimated that there over 400 million new cases of

DENV occur worldwide per year [110].

DENV affects several countries in South America and Asia and has recently started to affect Europe. Local transmission was observed in Europe for the first time in France and Croatia in 2010, placing Europe at risk of recurrent DENV epidemics [111]. In 2012 there was an outbreak of DENV in the Madeira islands of Portugal [112]. In the WHO Western Pacific Region, in 2017, over 140 thousand cases were reported [113]. In South America, in 2015, is estimated that there were 2.35 million cases of Dengue Fever [114]. In Africa, there is a probable under-recognition of DENV, although at least 22 African countries have reported to date sporadic cases or outbreaks since 1960 [109].

3.1.3 Yellow Fever virus transmission cycle and historical data

YFV was first isolated from a Ghanaian patient known as Asibi in 1927 [115]. It is believed that the virus originated a long time before that in Africa, and that it was spread from Africa into Europe and the Americas due to the slave trade between the continents [116]. Historical studies mention the presence of the first YFV epidemic in the Western Hemisphere, at the time known as "Black Vomit", as early as 1495 in the coastal areas of Central America [117]. YFV is a single serotype, although genotypes can be distinguished in Africa and South America [118].

YFV is transmitted to humans mainly by the *Aedes aegypti* mosquito. The transmission cycle can be divided into two parts: a natural endemic cycle and a secondary transmission to humans. The natural endemic cycle in the forest consists of a transmitting cycle between forest mosquitoes and wild primates. The secondary transmission to humans is subdivided into three cycles: sylvatic, intermediate and urban. The sylvatic cycle occurs in the tropical rain forests in Africa and South America. In this cycle, which is similar to the natural endemic cycle, occasionally the virus is transmitted from forest mosquitoes to humans, potentially creating patients "zero". The intermediate cycle is found in rural areas, where there is a more regular interaction be-

tween humans and primates, as well as the presence of semi domesticated mosquitoes as vectors, potentially generating small scale epidemics. The intermediate cycle is the most common kind of outbreak in Africa. The urban cycle, the major cause of concern for public health, occurs in high human population density areas with the presence of the urban mosquitoes *Aedes aegypti*, creating a mosquito-human-mosquito interaction that elevates the potential of spread of the virus considerably [119].

Currently, YFV is endemic to tropical regions in Africa and the Americas. As of 2016, 34 countries in Africa and 13 in Latin America are considered endemic for or have regions that are endemic for the YFV [120]. Recent outbreaks include an epidemic in Africa that started in Angola in 2015, with linked cases in the Democratic Republic of Congo, Uganda and China. A vaccination campaign in African countries led to the end of the outbreak in late 2016 to early 2017. The 2016/2017 epidemic in Brazil, is waning, but there have been over 723 confirmed cases including 237 deaths between July 2017 and February 2018 [121].

3.1.4 Zika virus transmission cycle and historical data

As mentioned in chapter one, ZIKV was first isolated in Africa [7]. Serological evidence shows the virus may have been present in Asia at the same period. Given that serological evidence at the time could not distinguish ZIKV from other close related viruses, the most accepted theory is that ZIKV originated in Africa.

Aedes aegypti is the main vector for the ZIKV, whereas *Aedes albopictus* plays a secondary role on the infection. Several other species of the genus *Aedes* are known to host ZIKV in Africa [45]. Diallo et al., 2014 [122] reports isolation of ZIKV from mosquitoes in the *Culex* genera, *Anopheles*, and *Mansonia* in Senegal. Recent report from Brazil have identified *Culex* species as a competent vector in the lab [123], although multiple groups claim *Culex* does not support replication of ZIKV [124, 125, 126].

The ZIKV transmission cycle is divided in sylvatic and urban cycles, with possibly

an intermediary cycle where there is an interaction between sylvatic and urban in rural areas. The sylvatic cycle consists of non-human primates and mosquito vectors. The urban cycle consists of the mosquito-human-mosquito interaction. Other mosquitoes also play a role as vectors in the sylvatic cycle of the disease but have minor impact on the urban transmission network as they are not present in the urban regions. There is also evidence of human-to-human sexual transmission [127, 128] thus making for potential global geographic spread of ZIKV.

Currently, ZIKV has been under control after an outbreak that started in 2015 in Brazil and culminated in a worldwide epidemic. A small number of cases are still being reported in Asia and the Americas [129].

Table 3.1: Epidemiological comparison of CHIKV, DENV, YFV and ZIKV.

Virus	**Vector**	**Isolation / First time seen**	**Transmission Cycle**	**Current Strains/Serotypes**
CHIKV	*Aedes spp.*	First isolated in Tanzania, 1952 / Sailing ships 18th century.	Sylvatic (non-human primates and mosquitoes) and Urban (mosquito-human-mosquito).	Asian lineage, Indian Ocean Lineage (IOL), Eastern Central and Southern African Lineage (ECSA) and West African.
DENV	*Aedes spp.*	First isolated in 1943/45 in Japan and Hawaii / Chin Dynasty between 265-420 AD.	Sylvatic (non-human primates and mosquitoes) and Urban (mosquito-human-mosquito).	DENV-1, DENV-2, DENV-3 and DENV-4.
YFV	*Haemagogus leococelaenus, Aedes spp.*	First isolated in Ghana, 1927 / Black Vomit in Central America, 1495	Natural Endemic; Secondary transmission to humans: Sylvatic (non-human primates, mosquitoes and humans), intermediate (rural areas) and Urban (mosquito-human-mosquito).	2 in South America / 5 in Africa (West Africa I, West Africa II, East and Central African, East African, Angola).
ZIKV	*Aedes spp.*	First isolated in Uganda, 1947 / Africa, unknown	Sylvatic (non-human primates and mosquitoes), Intermediary (rural areas), Urban (mosquito-human-mosquito) and Sexual transmission.	African / Asian / Asian-Pacific-American.

3.2 Material and Methods

All the geographic, sequence and temporal metadata for ZIKV and CHIKV previously generated were analyzed using the pipeline described below (Figure 3.1). Also, datasets were created for Yellow Fever and all four Dengue serotypes, utilizing all data available in the public domain (NCBI) as of April 1st 2018. Selection for the appropriate outgroup for Dengue and Yellow Fever viruses will be made based on the phylogeny of *Flavivirus*. To investigate the phylogenetic relationships, I used the maximum likelihood tree search method as implemented in IQ-TREE [55].

The transmission network pipeline utilizes the following steps:

1 - Sequence data is acquired from GenBank; 2 - Datasets are built with genomic sequences and metadata associated; 3 - Multiple sequence alignments are performed with MAFFT; 4 - Maximum Likelihood tree search is performed with IQ-Tree software; 5 - Ancestor-descent changes in metadata states are calculated within R script; 6 - The transmission network is built based on ancestor-descendent changes; 7 - Centrality measurements and a newly introduced Source/Hub Ratio are calculated on the network to identify important hubs and source of spread of the disease.

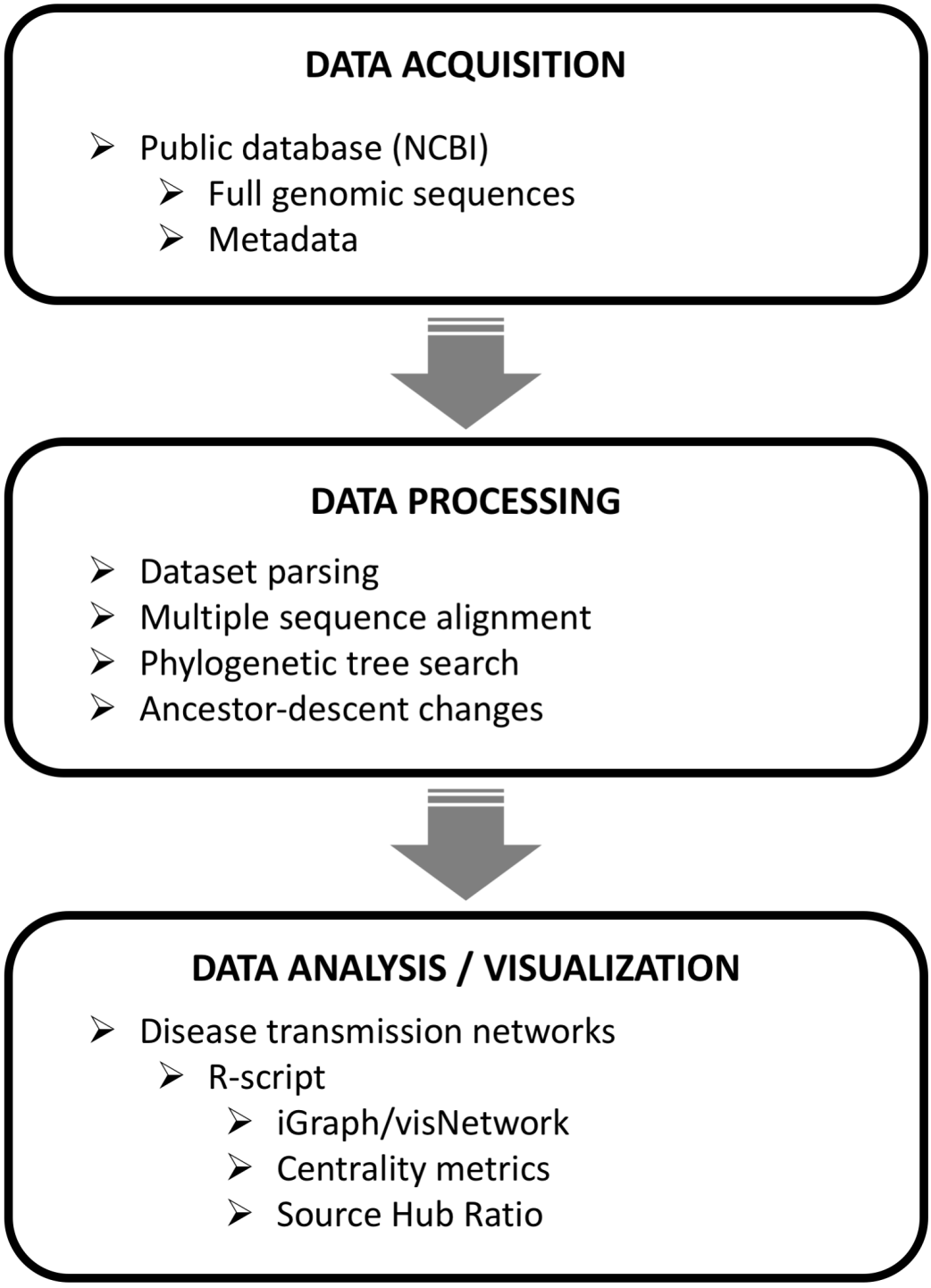

Figure 3.1: Transmission Network Pipeline.

3.2.1 Datasets

Datasets were generated comprising all full genomic sequence available for Zika, Chikungunya, Dengue 1, Dengue 2, Dengue 3 and Dengue 4 and Yellow Fever virus.

The creation of the datasets follow the steps below:

1. Access the public database and download all fasta files and gb files (full NCBI sequence with metadata associated);

2. All metadata files will be parsed and isolates with >5-fold passage history sequences will be removed;

3. Create the fasta file with filtered sequences and spreadsheet with ID and metadata of interest from gb file.

3.2.2 Multiple Sequence Analyses

Multiple sequence alignment were performed as described on Chapter 2.

3.2.3 Phylogenetic Tree Search

Phylogenetic tree search were conducted as described on Chapter 2.

3.2.4 Ancestor descent Changes

To obtain the ancestor descent changes (also known as an apomorphy list), the phylogenetic tree generated on the phylogenetic tree search step was merged with the metadata in nexus format on Mesquite. The metadata was traced to the tree using the parsimony ancestral state reconstruction method for validation of the file. Once validated, the nexus file was run on R script utilizing the package 'castor', which calculates the state changes and creates an apomorphy list.

3.2.5 Transmission Networks

The transmission network was created using data extracted from ancestor descent changes and visualized utilizing R packages igraph and visNetwork. Multiple estab-

lished centrality measurements and novel metrics (i.e. Source Hub Ratio (SHR)) described below were calculated to evaluate the transmission networks (Table 3.2).

3.2.6 Centrality Measurements

Multiple centrality measurements are available to determine the relative importance of nodes within a network. In this work, the transmission network nodes were evaluated using three different centrality metrics: betweenness, closeness and degree.

Degree centrality is the number of links a node has with other nodes within the network. The higher the number of the links would indicate more importance in terms of centrality. Betweenness centrality measures the centrality of a node given the number of shortest paths between two other nodes that passes through the node of interest, normalized by all pairs of node within the network. Closeness centrality measures the centrality of a node based on the relative sum of all the shortest paths from that node to all other nodes within network [130].

3.2.7 Source/Hub Ratio

The Source Hub Ratio (SHR) calculates the importance on a node within the network as source of the disease, ignoring centrality. It utilizes the concept of indegree and outdegree from directed networks, in terms of how many ties are generated from and to the node and calculates a value ranging from 0-1 that reflects the importance of the node as the source, from where the disease in other locations originates, or hub, where there is a similar amount of transmission of the disease going in and out of the location. Nodes with values close to 0.5 are equal to a hub for the disease, 1 equal to source of disease, and 0 equals to a dead end from which the disease does not spread. The SHR of a node "i" is equals to the sum of all shifts from location "i" to other locations ($\sum$ source(i)), normalized over the sum of all shifts from and to location "i" ($\sum$ hub(i)). This metric provides a notion of how important a node is within the network as source of the disease, ignoring centrality within network.

Table 3.2: Metrics for evaluating disease transmission networks.

Metric	**Formula**	**Formula Explained**	**Epidemiological Meaning**
Betweenness Centrality	$C_B(i) = \sum \frac{gjk(i)}{gjk}$	Betweenness centrality of node i is the sum of all the ratios between all the paths g from node j to node k that go through node i over all possible paths from node j to node k.	How important a location is as the shortest intermediary location connecting other locations within the transmission network.
Closeness Centrality	$C_C(p_k) = \frac{n-1}{\sum d(p_i, p_k)}$	Closeness centrality of node k (pk) is the inverse of the average network distance between pk and the other nodes (pi).	How important a location/host is given its distance within the transmission network to other locations/hosts.
Degree Centrality	$C_D(p_k) = \frac{\sum a(p_i, p_k)}{n-1}$	Degree centrality of node k (pk) is the sum of all links of other nodes (pi) to node k over the total number of nodes -1.	How important a location is within the transmission network given the number of times that a disease emerges from or to that point (indegree or outdegree).

Table 3.2 continued from previous page

Metric	Formula	Formula Explained	Epidemiological Meaning
Source/Hub Ratio	$SHR(i) = \frac{\sum source(i))}{\sum hub(i)}$	The Source Hub Ratio of a node i is the sum of all shifts from location i to other locations over the sum of all shifts from and to location i.	How important a node is within the network as source of the disease, ignoring centrality within network.

3.3 Results

3.3.1 Viral Transmission Networks

In order to understand the historical transmission dynamics of CHIKV, DENV1, DENV2, DENV3, DENV4, YFV and ZIKV, individual transmission networks were created utilizing phylogenetic trees containing all the full genomic sequences available for the viruses on an in-house R script. The node sizes were adjusted by betweenness, closeness, degree centrality metrics as well as SHR. Betweenness centrality was selected as the best metric to scale the nodes in large networks as it enlarges the nodes within the network that may serve as the shortest path between nodes, thus acting like a hub for the spread of disease (Table 3.3).

On less complex networks, SHR seems to be good to scale the nodes, as betweenness centrality doesn't bring much information to the network graph. Nevertheless, this is only true for less complex networks, as the SHR scale is not a linear increase, with values close to 0 representing dead-end, .5 representing hubs and close to 1 representing only the source of virus, making difficult to distinguish the importance of the nodes within a large network by simply eyeballing without the actual numbers. Thus the reason why I only used the SHR to scale nodes on the YFV Transmission Network graph as a supplement of information to the Betweenness Centrality graph. YFV was the smallest dataset with only 147 sequences.

3.3.1.1 Chikungunya virus

The historical transmission network of CHIKV was generated from a tree with 693 full genomic sequences (Figure 3.2). The results show that with the current available data, despite of the recent outbreaks in the Americas, the largest hub of the virus is Southeastern Asia, followed by Southern Asia. In Africa, the main hub is Eastern Africa and in the Americas is Central America. Not surprisingly, by comparing these findings with the CHIKV phylogenetic tree topology (Figure 2.2), the Asian hubs

match with the locations where the Asian Urban and IOL lineages are endemic. In Africa, Eastern Africa is the main hub, and was the origin of the IOL and in the Americas, Central America is the largest as it spread from the Asian Urban lineage. I built subtrees for CHIKV to investigate in depth the dynamics of its multiple lineages, the data is presented below.

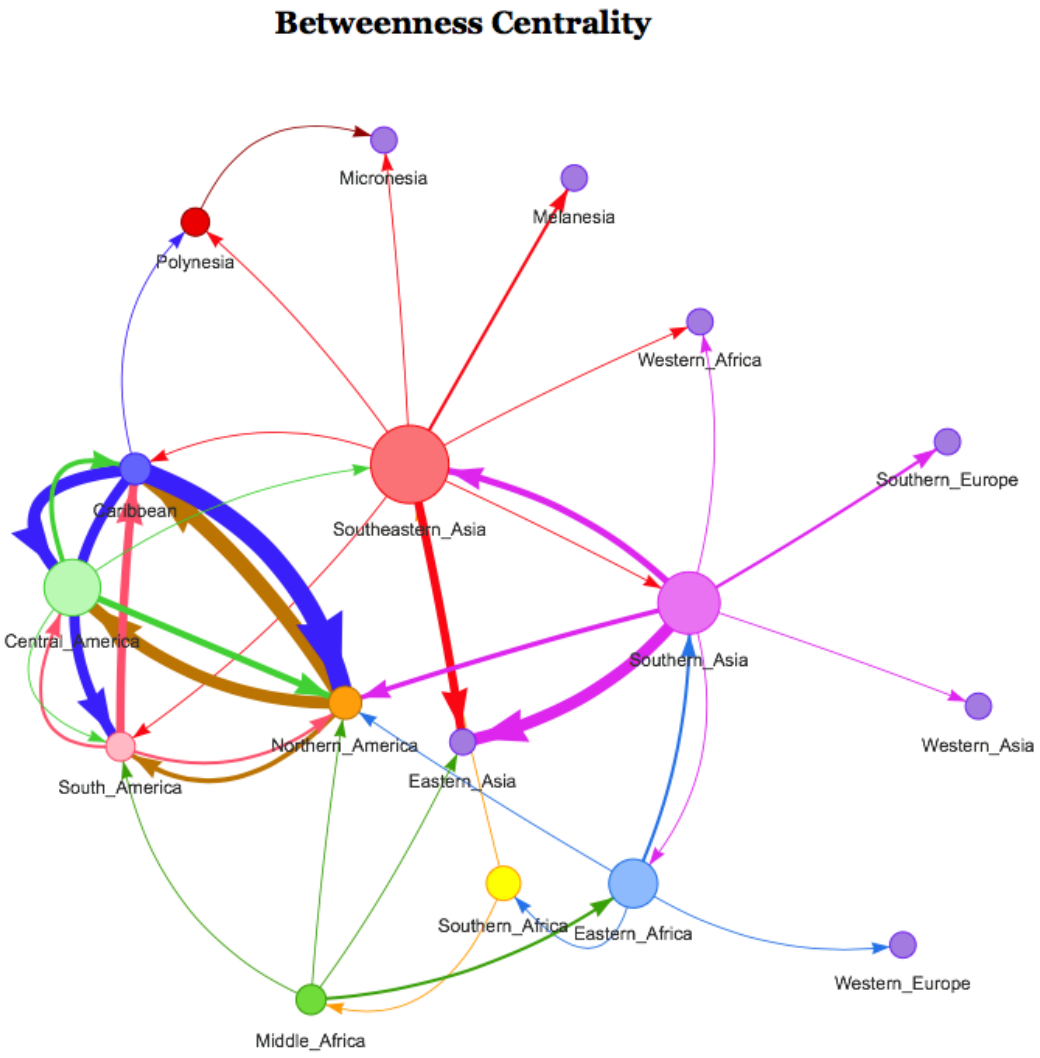

Figure 3.2: Transmission Network of Chikungunya virus. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

3.3.1.2 Dengue virus

The historical transmission network of DENV was divided in four networks, one for each serotype. Given that the genetic similarity between the serotypes is similar to when comparing different species within flaviviruses, it would make no sense in combining the four datasets into a master dataset as one would expect that the serotypes are evolving independently. DENV-1 is the largest dataset with 2106 sequences, DENV-2 has 1535 sequences, DENV-3 1039 and DENV-4 356 sequences.

DENV-1 has the largest hubs in Eastern, South and Southeastern Asia in the Asian Continent, and Northern and South America in the Americas, surprisingly, not a large betweenness centrality is found within Central America and Caribbean (Figure 3.3). In contrast, DENV-2 has the largest hubs in the Caribbean and South America in the Americas, while in Asia has the largest hub in Southeastern Asia (Figure 3.4). DENV-3 has its main hub in South America in the Americas and Eastern Asia in Asia (Figure 3.5). DENV-4 has the largest hub in Southeastern Asia in Asia and South America in the Americas.

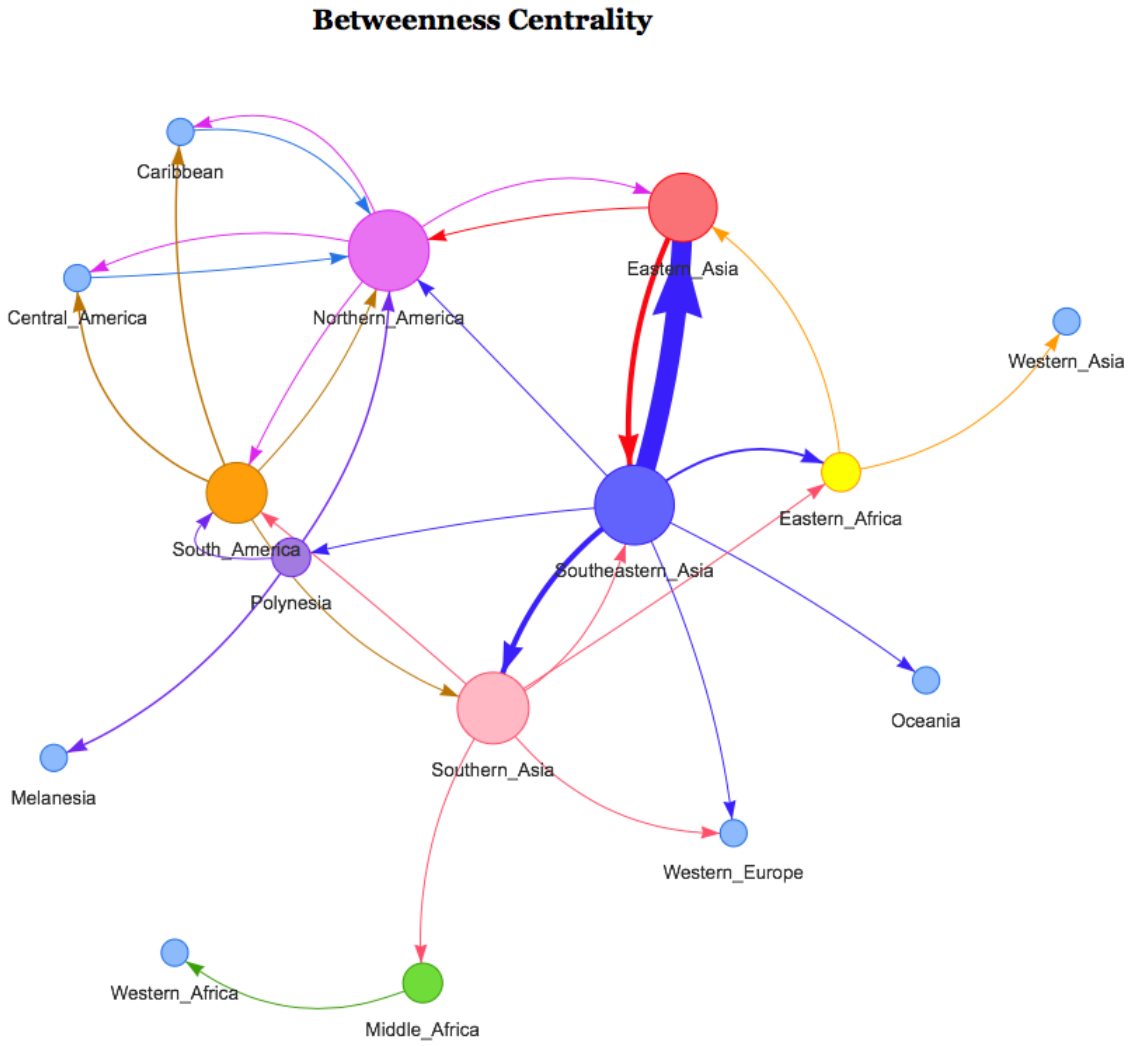

Figure 3.3: Transmission Network of Dengue virus serotype 1. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

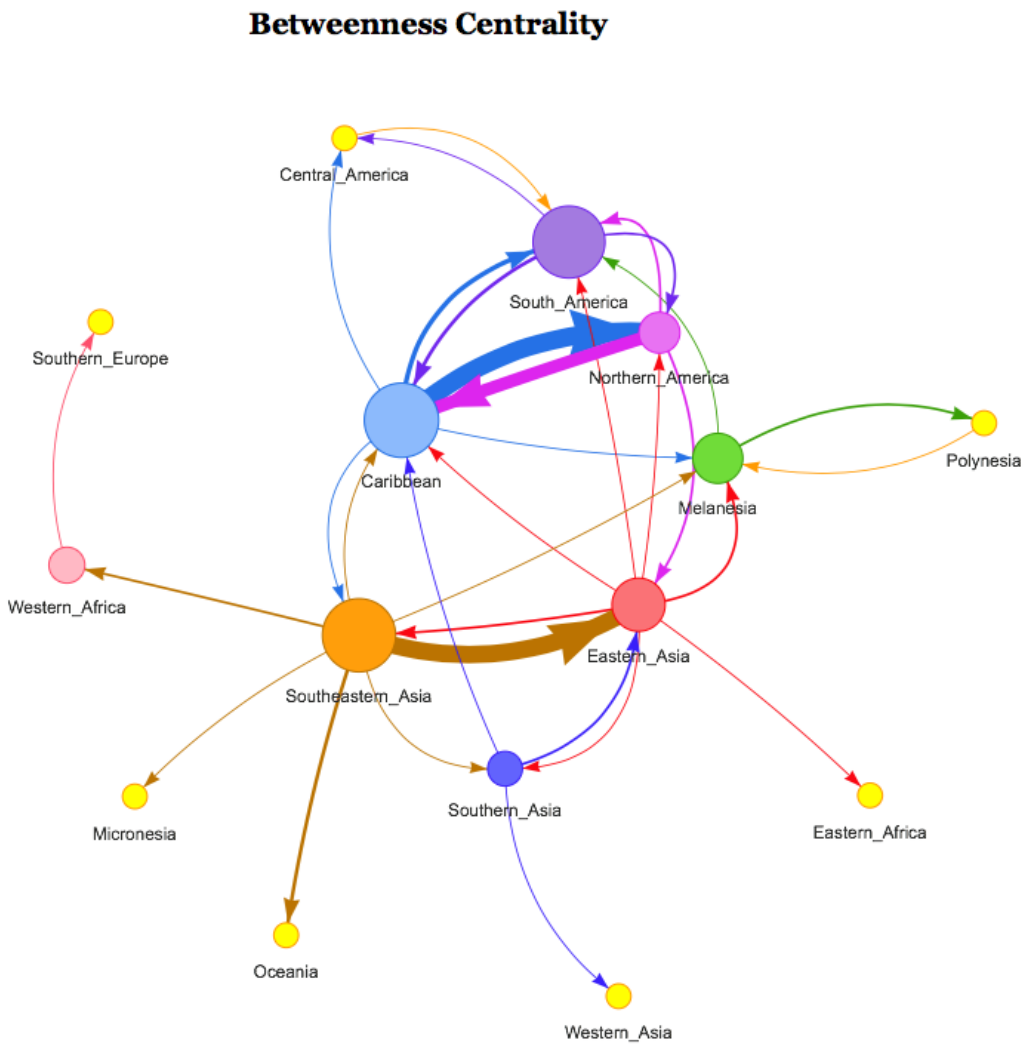

Figure 3.4: Transmission Network of Dengue virus serotype 2. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

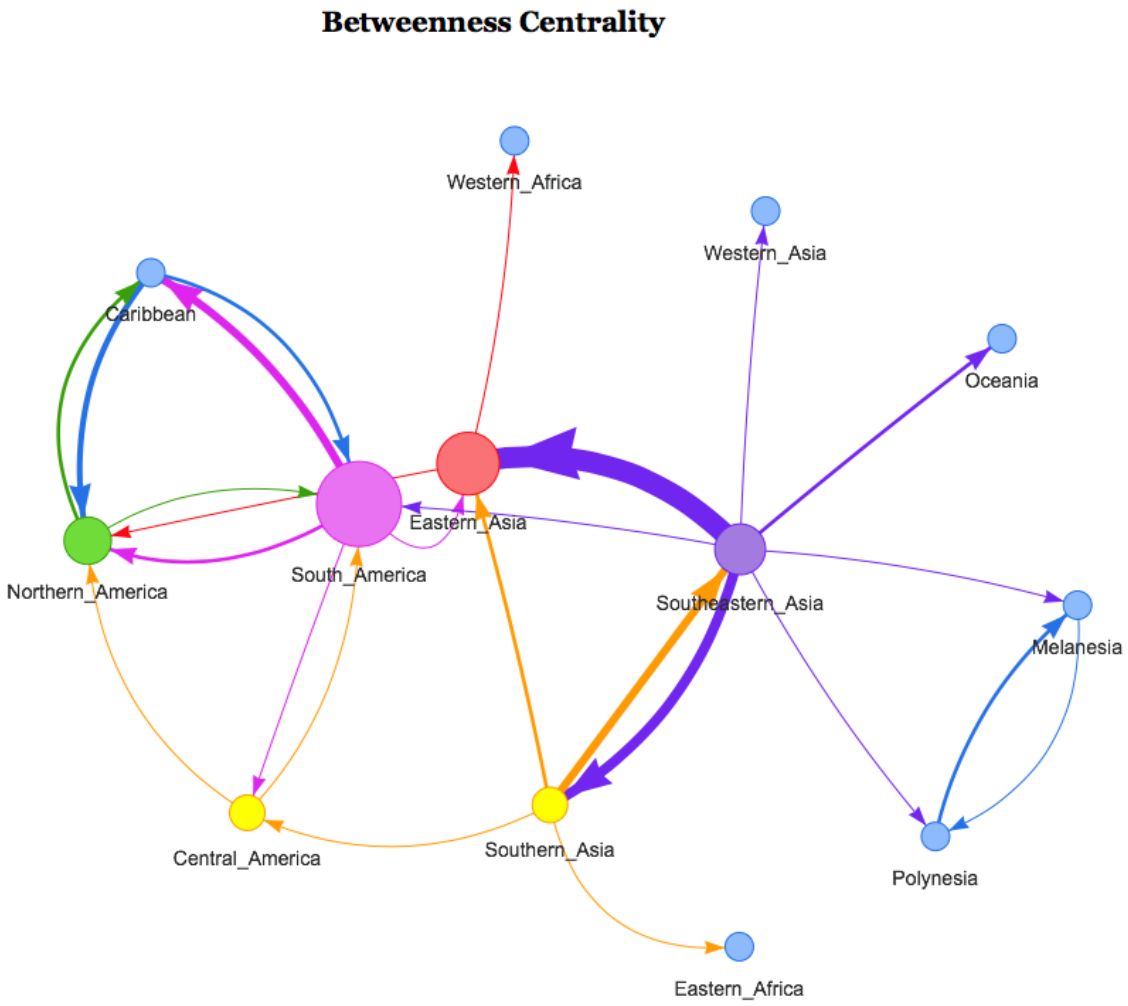

Figure 3.5: Transmission Network of Dengue virus serotype 3. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

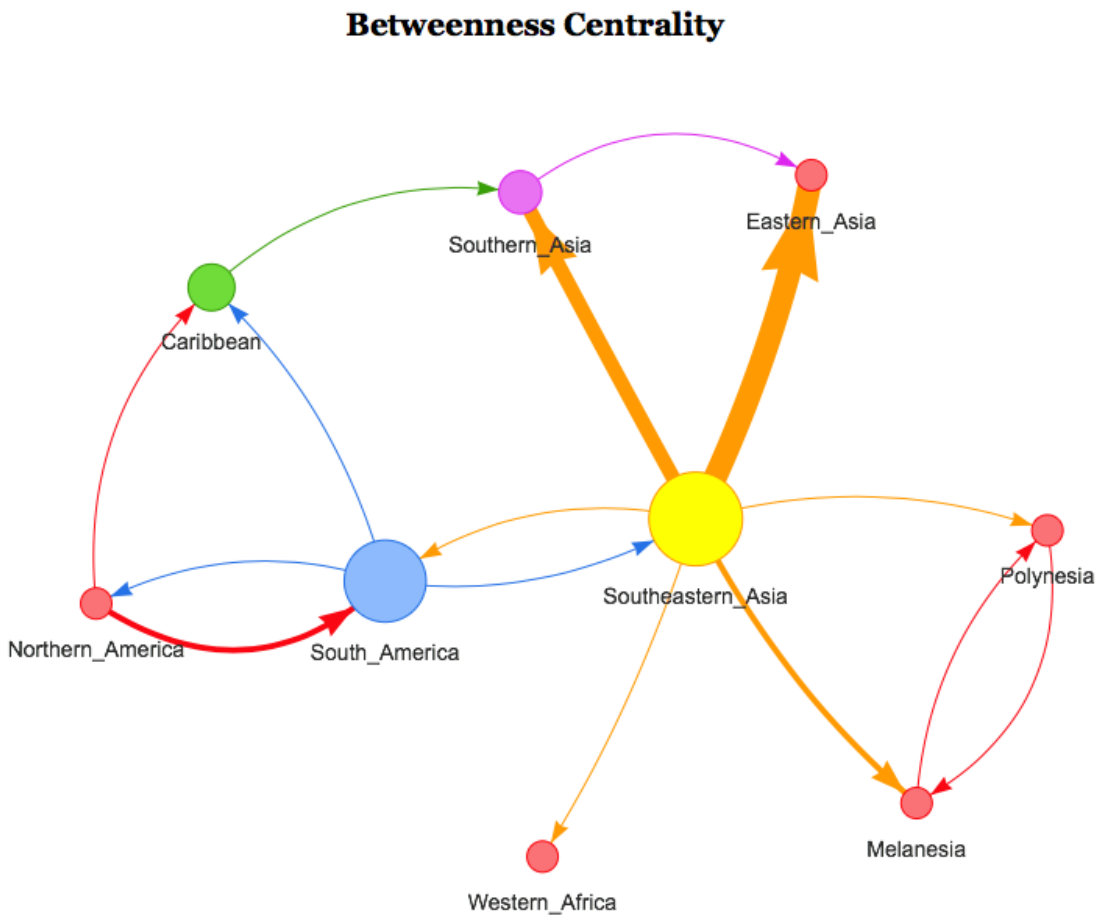

Figure 3.6: Transmission Network of Dengue virus serotype 4. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

3.3.1.3 Yellow Fever virus

As mentioned previously, YFV was the dataset with the lowest number of sequences, 147, consequently, the number of transmission events were smaller and generated the less complex network over all networks generated on this work. Western Africa is among the geographic regions with the lowest betweenness centrality score, while Middle Africa has the highest score, followed by South America (Figure 3.7). Although betweenness centrality seems to work well on larger networks, it seemed

that on smaller networks due to the lack of available data centrality metrics do not provide all the information someone would like to infer about the network. In the CHIKV transmission network, SHR scaled the nodes in a way that it made it easy to discern Western Africa (SHR = 1) from other nodes as the major source of the disease and Middle Africa (SHR = 0.6), South America (SHR = 0.5), Eastern Africa (SHR = 0.5) and Eastern Asia (SHR = 0.5) act as hubs (Figure 3.8). By associating the information between the betweenness centrality metric and SHR, we can assume that the main hubs in terms of centrality within the network are South America and Middle Africa, although we can also note that Eastern Asia and Eastern Africa have multiple shifts of character states that could potentially justify them as secondary hubs to be investigated.

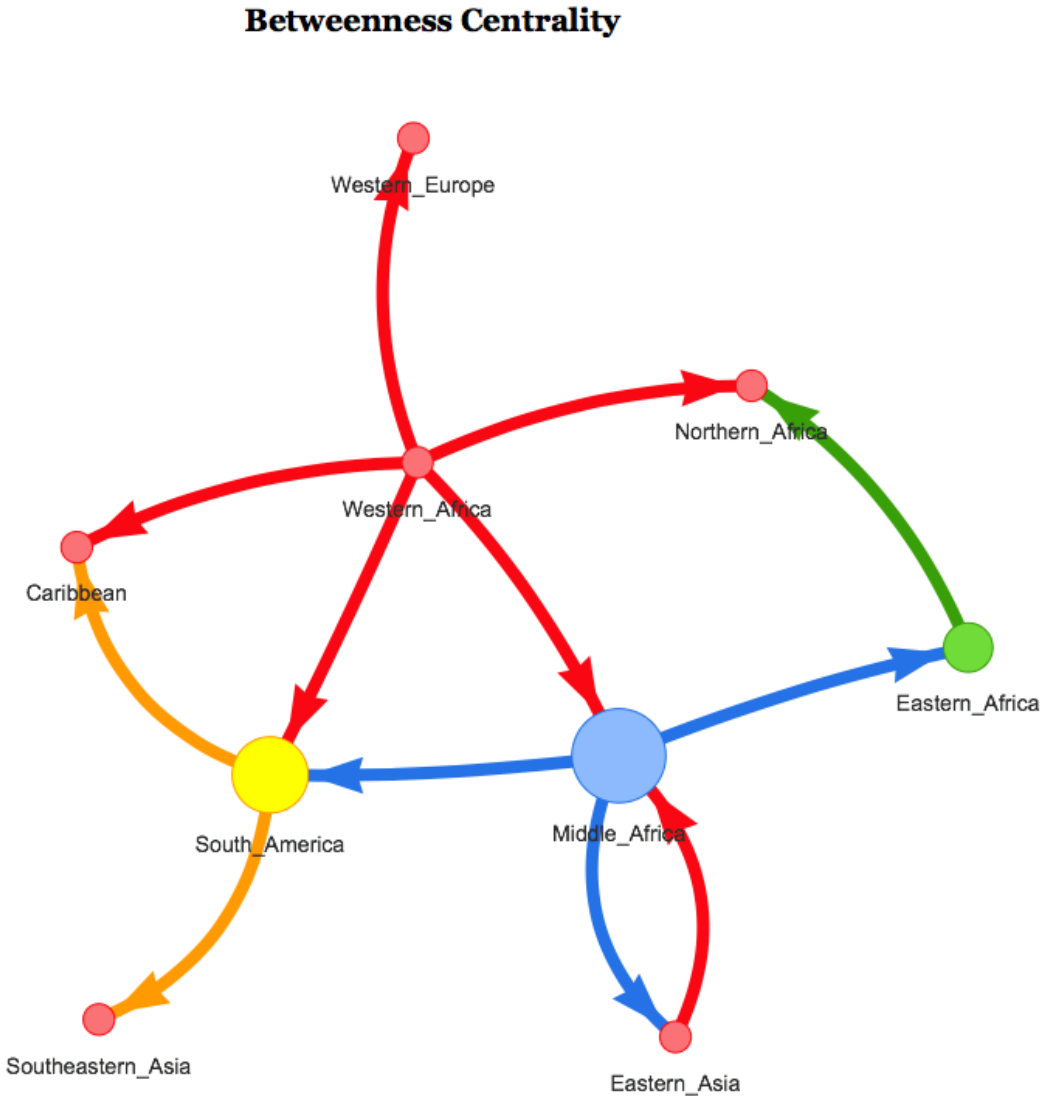

Figure 3.7: Transmission Network of Yellow Fever virus. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

Figure 3.8: Transmission Network of Yellow Fever virus. Nodes represent geographical regions according to UN Geoscheme and scaled by Source Hub Ratio metric (larger = more SHR). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

3.3.1.4 Zika virus

The historical transmission network of ZIKV was generated from a tree with 490 full genomic sequences (Figure 3.9). The largest hub is in Eastern and Southeastern Asia

in the Asian continent. In Africa, Eastern Africa is the main hub and in the Americas the Caribbean is the main hub, followed by South America. Using as reference the ZIKV phylogenetic tree topology (Figure 2.3), is interesting to see the role of Eastern and Southeastern Asia as a hub within the historical transmission network of ZIKV, given that for exception of the Asian lineage it seemed to evolve linearly with a larger movement within the Americas. Along with CHIKV, I built subtrees for ZIKV to investigate in depth the dynamics of these two viruses with multiple lineages, which are presented below.

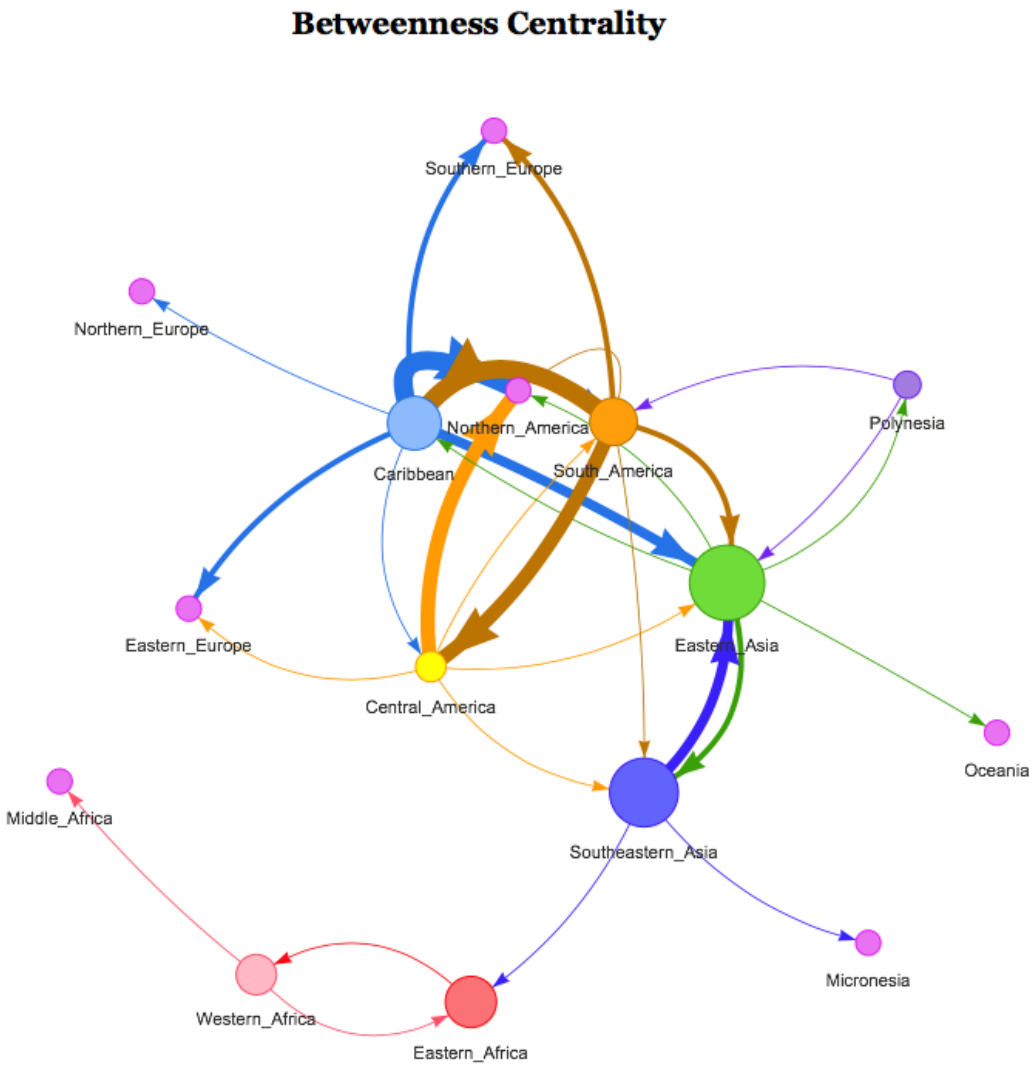

Figure 3.9: Transmission Network of Zika virus. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

Table 3.3: Summary of historical transmission networks results.

Virus	Main disease spread hubs
Chikungunya	Eastern Asia; Southeastern Asia; Caribbean
Dengue 1	Southeastern Asia; Northern America; Southern Asia; Eastern Asia
Dengue 2	Southeastern Asia; Caribbean; South America;
Dengue 3	South America; Eastern Asia; Southeastern Asia; Northern America
Dengue 4	Southeastern Asia; South America
Yellow Fever	Betweenness Centrality: Middle Africa; South America / SHR: Middle Africa; Eastern Africa; South America; Eastern Asia
Zika	Southeastern Asia; Southern Asia; Eastern Africa; Central America

3.3.2 Specific Lineage Transmission Networks

While it is interesting to include the maximum amount of genomic data available from a pathogen to build phylogenetic relationships, building transmission networks out of these large trees do not provide insights to specific events or outbreaks, but rather than this it gives an overall look on the behavior of the disease over the years, an historical overview. Although it can introduce new insights on the overall documented history of the virus, mixing the data from different major clades/lineages can mask the real origin of specific outbreaks. In order to investigate in depth the major players on the history of separate CHIKV and ZIKV outbreaks, I created multiple subtrees based on major clades of interest and generated individual transmission networks with nodes scaled by Betweenness Centrality at two levels of granularity, UN Geoscheme and Country. The summary of my findings can be found on Table 3.4.

3.3.2.1 Chikungunya virus

The transmission networks of CHIKV were generated from four subtrees from the original 693 full genomic sequences dataset. The subtrees were labeled CHIKV Asian Urban Lineage, CHIKV Indian Ocean Lineage, CHIKV South American and CHIKV

West African Lineage which encompasses the American Lineage and Asian Urban Lineage, the Indian Ocean Lineage and Eastern African Lineage, the Middle Africa and South American Lineages, and Western African lineage, from Figure 2.2, respectively.

CHIKV Asian Urban dataset was formed with 411 genomic sequences. The largest hub within this network is in the Caribbean, with Polynesia also being a secondary hub in Asia (Figure 3.10). Expanding the network by looking in depth into the shifts between countries, the Caribbean Islands remain the largest hub within America, and the Philippines are a small but secondary hub in Asia (Figure 3.11). When looking on the topology of the subsection of tree which represents this dataset (Figure 2.7) it can be observed that there is not a single Asian Urban Lineage clade, but multiple small Asian clades until this lineage reaches the Americas, forming an "American Lineage", which may justify why there is a node with large betweenness centrality in the Americas but not in Asia, the Asian sequences don't seem to be highly interconnected.

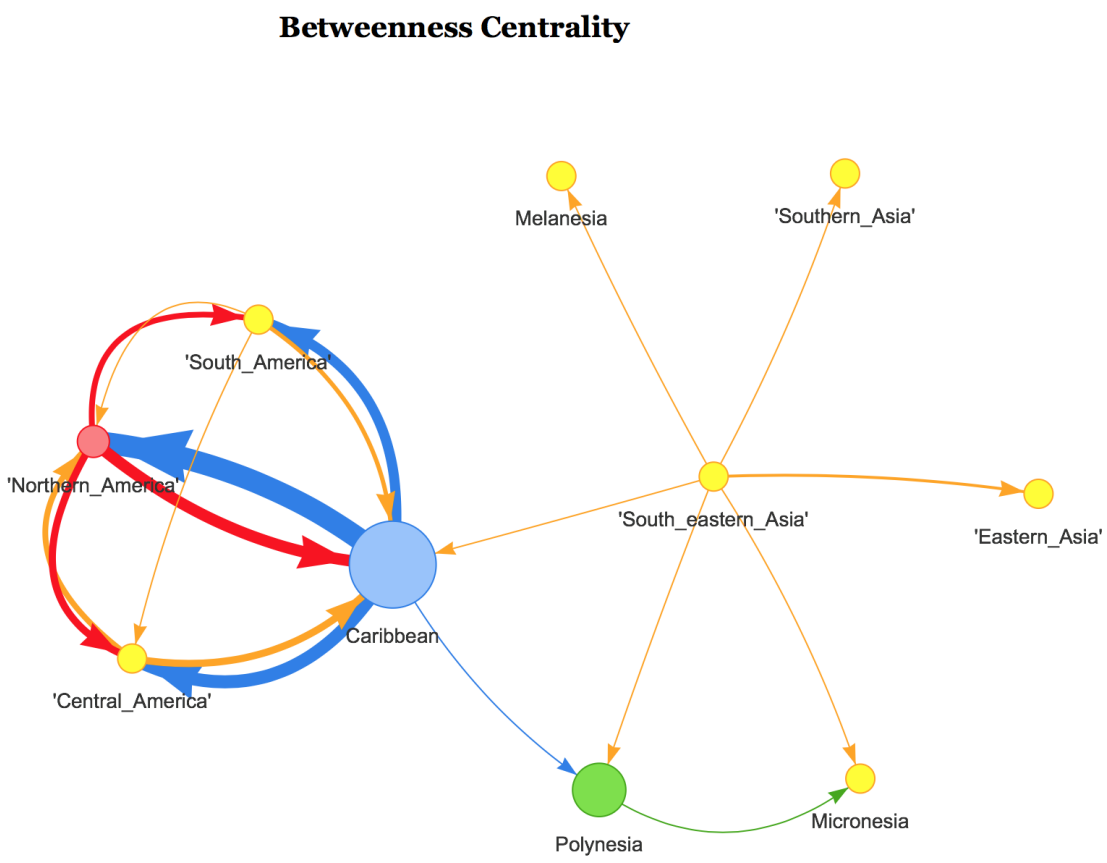

Figure 3.10: Transmission Network of Chikungunya virus Asian Urban lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

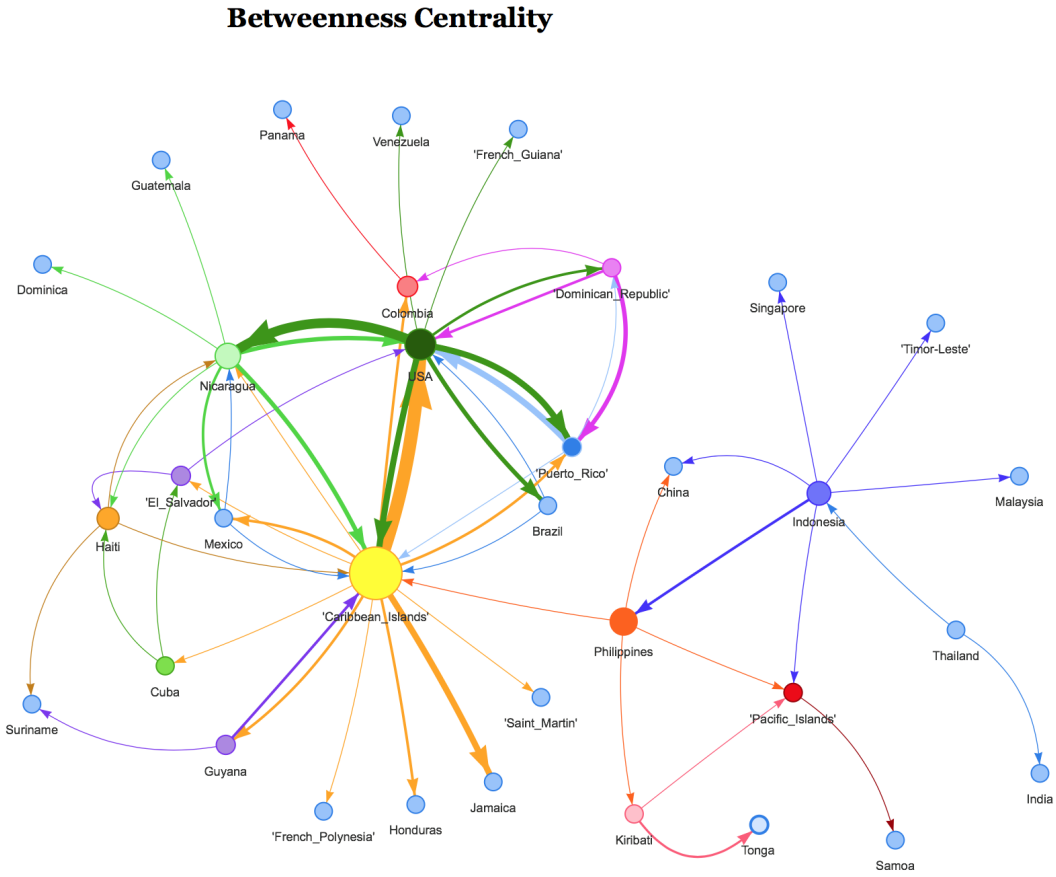

Figure 3.11: Transmission Network of Chikungunya virus Asian Urban lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

CHIKV Indian Ocean Lineage dataset was formed with 210 genomic sequences. The largest hub within this network is in Southern Asia, with Southeastern Asia as a secondary hub (Figure 3.12). Looking in depth within this network, Malaysia, Thailand and India are the main hubs within the network (Figure 3.13). These results show a difference in behavior when in comparison with the Asian Urban Lineage and can be easily explained when the phylogeny of the Indian Ocean Lineage is present (Figure 2.6) as the Asian sequences for this lineage form a monophyletic clade.

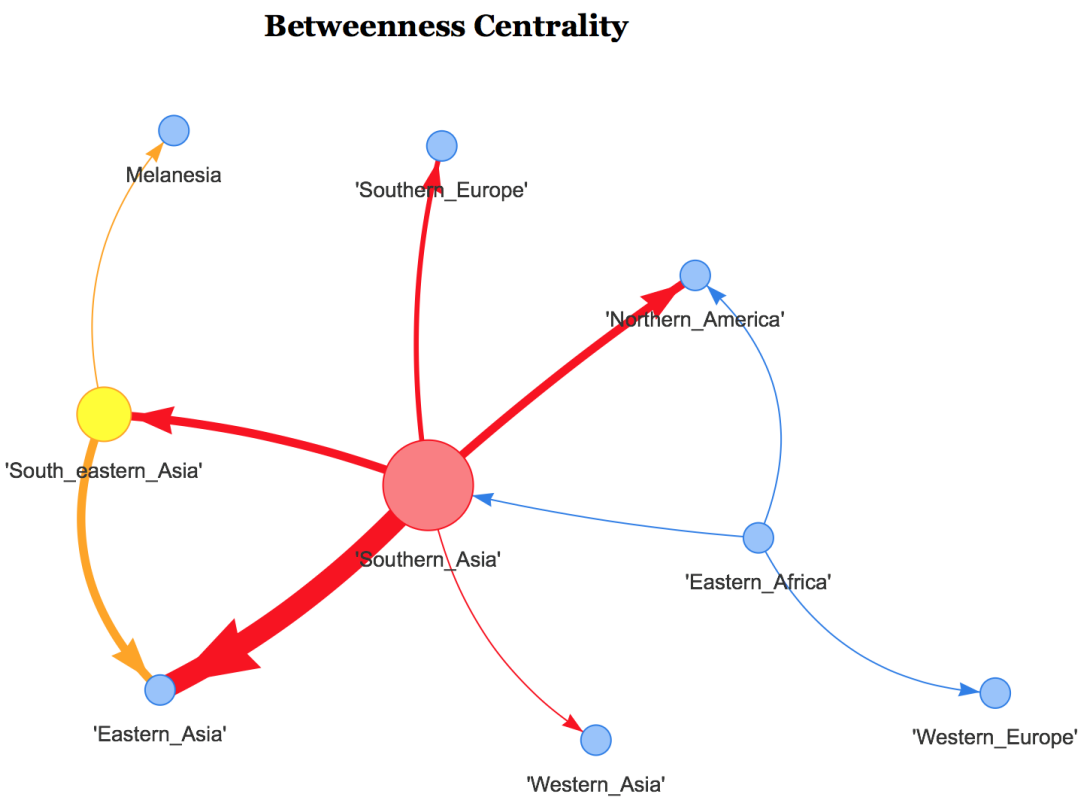

Figure 3.12: Transmission Network of Chikungunya virus Indian Ocean lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

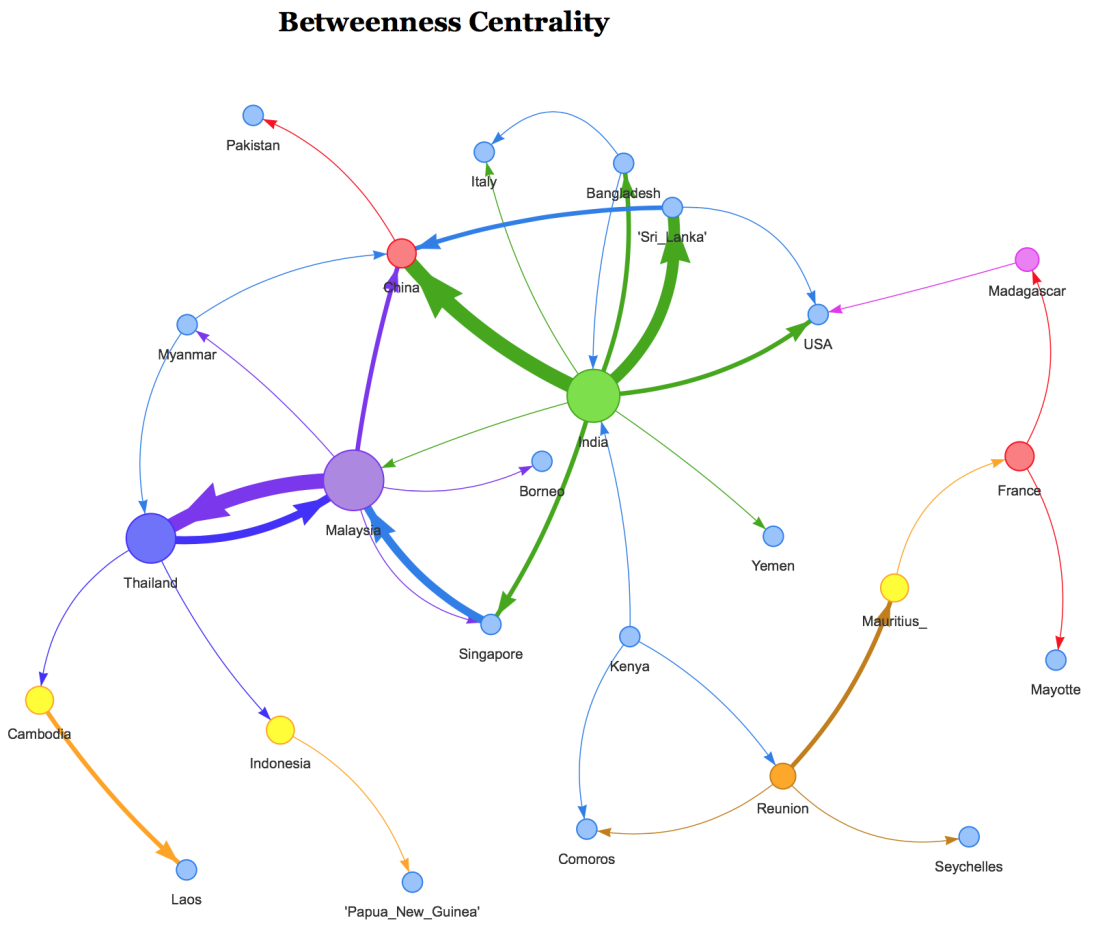

Figure 3.13: Transmission Network of Chikungunya virus Indian Ocean lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

CHIKV South American dataset was formed with 56 genomic sequences. Interestingly, the largest hubs for this outbreak were not South America nor Middle Africa, where one would expect hubs to be given the large amount of Middle Africa sequences and their positioning on the phylogenetic tree (Figure 2.5). Eastern Africa and Southern Asia, places of origin of the Indian Ocean Lineage, are the main hubs according to the built network (Figure 3.14). This result may be explained due to the increased effect of noise data (sister clades) on the shape of a network with small amount of genomic sequences.

Thus, I scaled the transmission network based on the SHR in order to clarify the role of Middle Africa, as well as Southern Asia and Eastern Africa, given that Source Hub Ratio doesn't account for centrality of the network. Middle Africa (SHR = 1) received a source score, which helps to explain why it was not flagged as a hub on the Betweenness Centrality measurement. Southern Asia (SHR = 0.67) and Eastern Africa (SHR = 0.33), received both a score close to 0.5, representing hubs and confirming the results from betweenness centrality (Figure 3.15). South America (SHR = 0.5) and Southern Africa (SHR = 0.5) also scored as hubs on SHR, but given that they are not central within the network, they were not considered hubs on betweenness centrality. Looking in depth at the country transmission graph, we can see the main hub within Africa was in Angola, and within Asia was in India (Figure 3.16). Uganda and Tanzania are the only two nodes within this network that represent Eastern African countries, together, they shared the highest betweenness centrality with Southern Asia, but once evaluated individually, they did not have a role as hubs within the network, which raises the importance of looking at iterations at high resolution on networks with low amount of data.

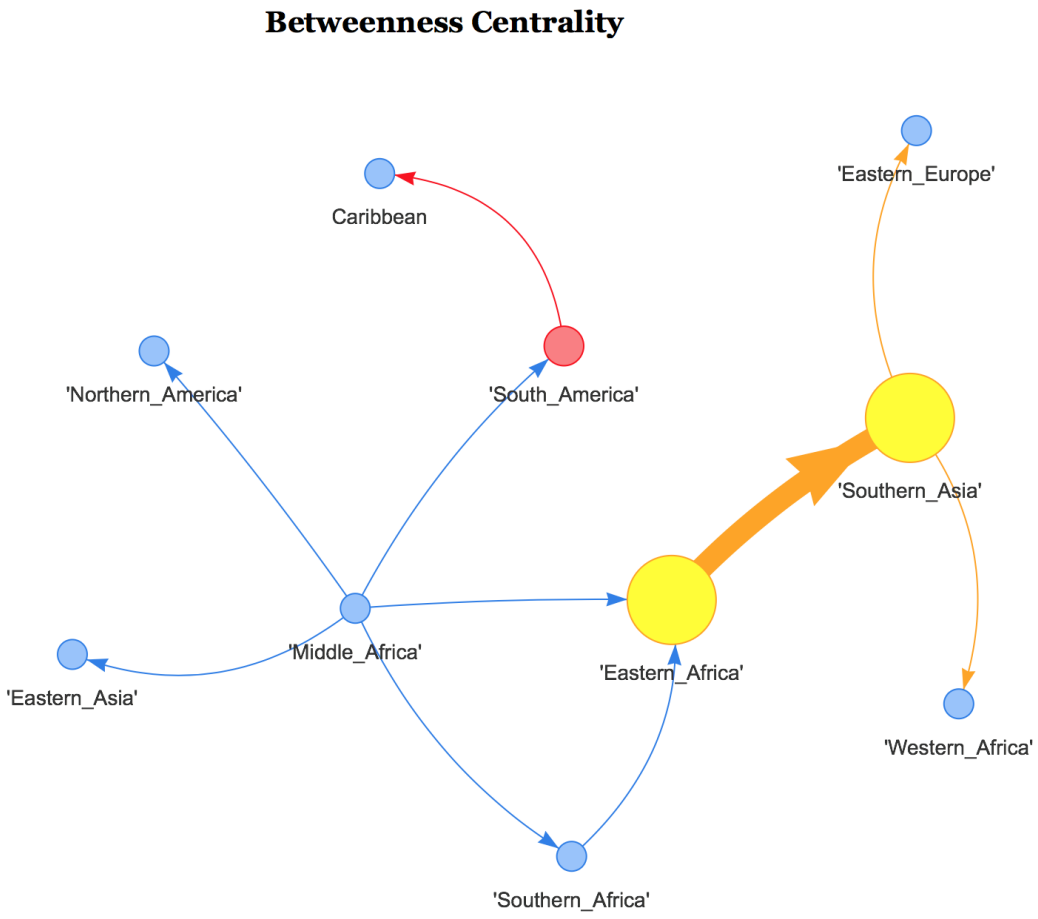

Figure 3.14: Transmission Network of Chikungunya virus South American lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

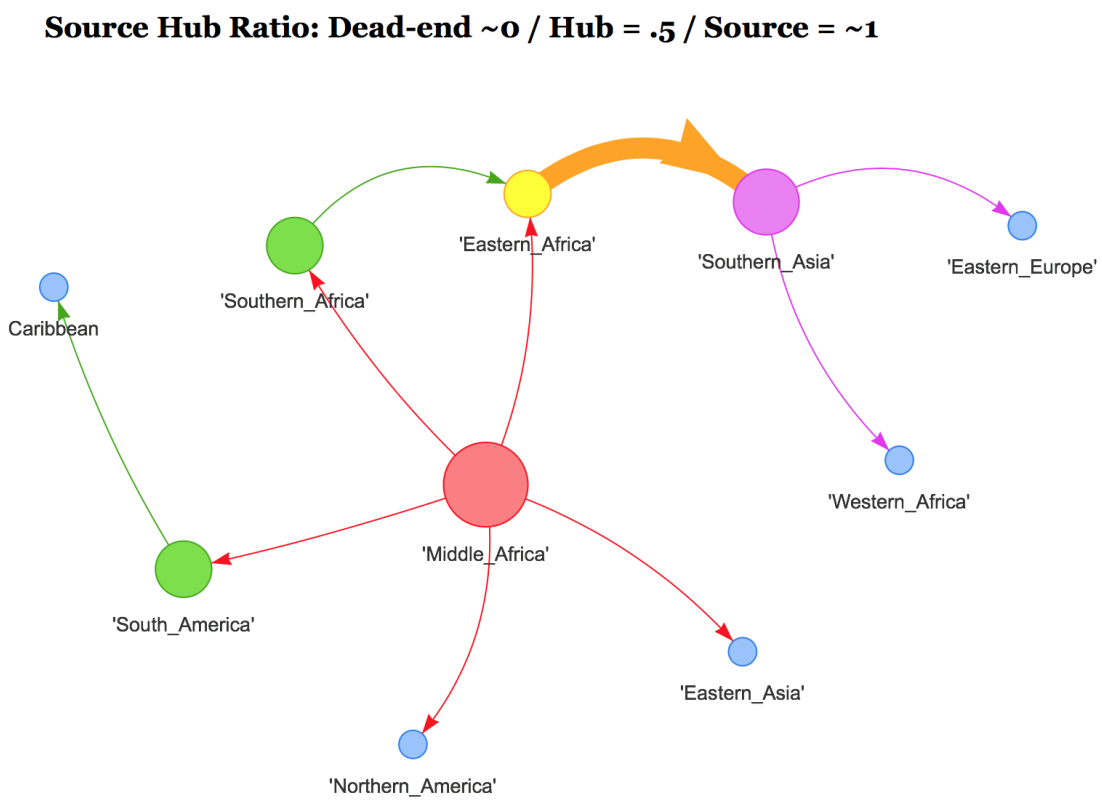

Figure 3.15: Transmission Network of Chikungunya virus South American lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by Source Hub Ratio metric (larger = more SHR). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

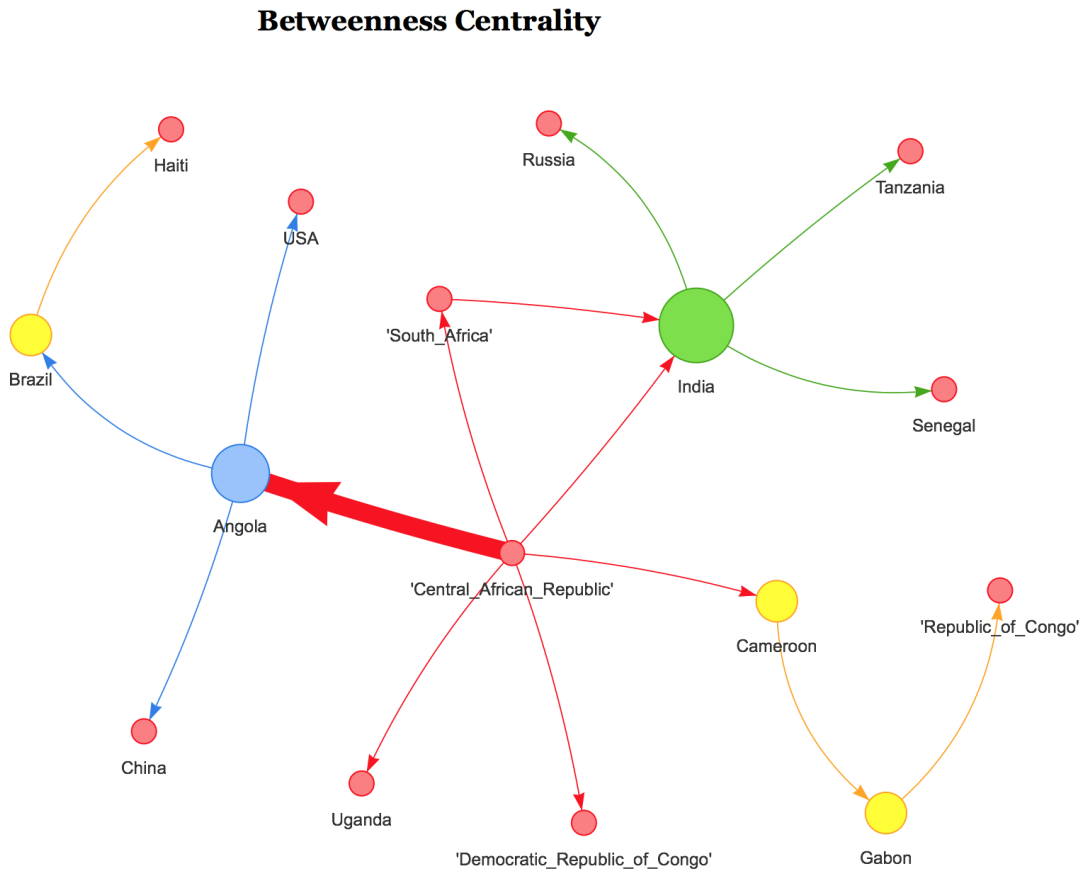

Figure 3.16: Transmission Network of Chikungunya virus South American lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

CHIKV West African dataset was formed with only 14 genomic sequences. Given the fact that all sequences within the West African Lineage haven't been found outside this region, the transmission network was only evaluated at "country" level. In this scenario, the only metric applied on the nodes to make sense was SHR given the small number of nodes (Figure 3.17). According to the available data, we can observe Cote d'Ivoire (SHR = 0.5) with shifts in and out to Senegal (SHR = 0.67), while Nigeria (SHR = 0) only had shifts originating from Senegal.

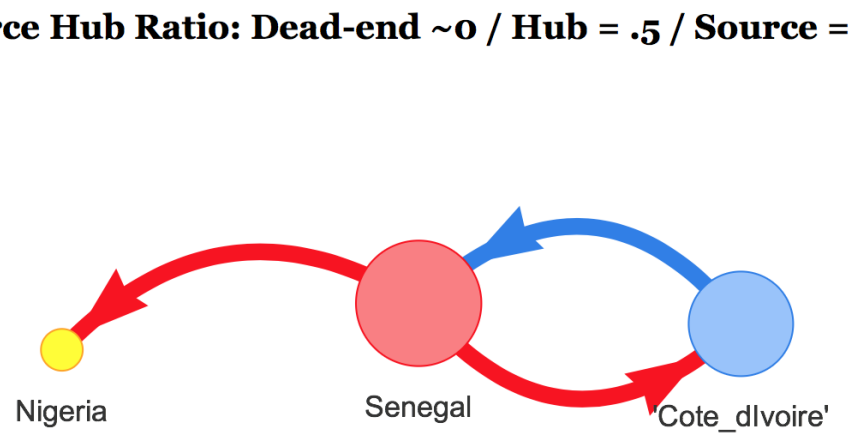

Figure 3.17: Transmission Network of Chikungunya virus West African lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

3.3.2.2 Zika virus

The transmission networks of ZIKV were generated from two subtrees from the original 490 full genomic sequences dataset. The subtrees were labeled ZIKV Asia Pacific American lineage and ZIKV African strains and Asian Lineage which encompasses the sequences from the Pacific Islands until the recent outbreaks in the Americas, and the old African sequences and Asian sequences until the recent outbreaks in Singapore and other cases found in Asia that formed the Asian Lineage, from Figure 2.3, respectively.

ZIKV Asia Pacific American dataset was formed with 325 genomic sequences. The largest hub in Asia is found in Eastern Asia, while in the Americas is in the Caribbean followed by South America (Figure 3.18). By calculating the Source Hub Ratio (Transmission Network not shown), Caribbean (SHR = 0.78), Central America (SHR = 0.71) and South America (SHR = 0.67) in a decreasing order had a role of Source and Hub for the disease in the Americas, while Polynesia (SHR = 0.66)

and Eastern Asia (SHR = 0.44) were hubs. By looking in depth on the transmission network, Dominican Republic, Brazil, Colombia and Mexico are the major hubs for this clade in the Americas, whereas in Asia is China (Figure 3.19).

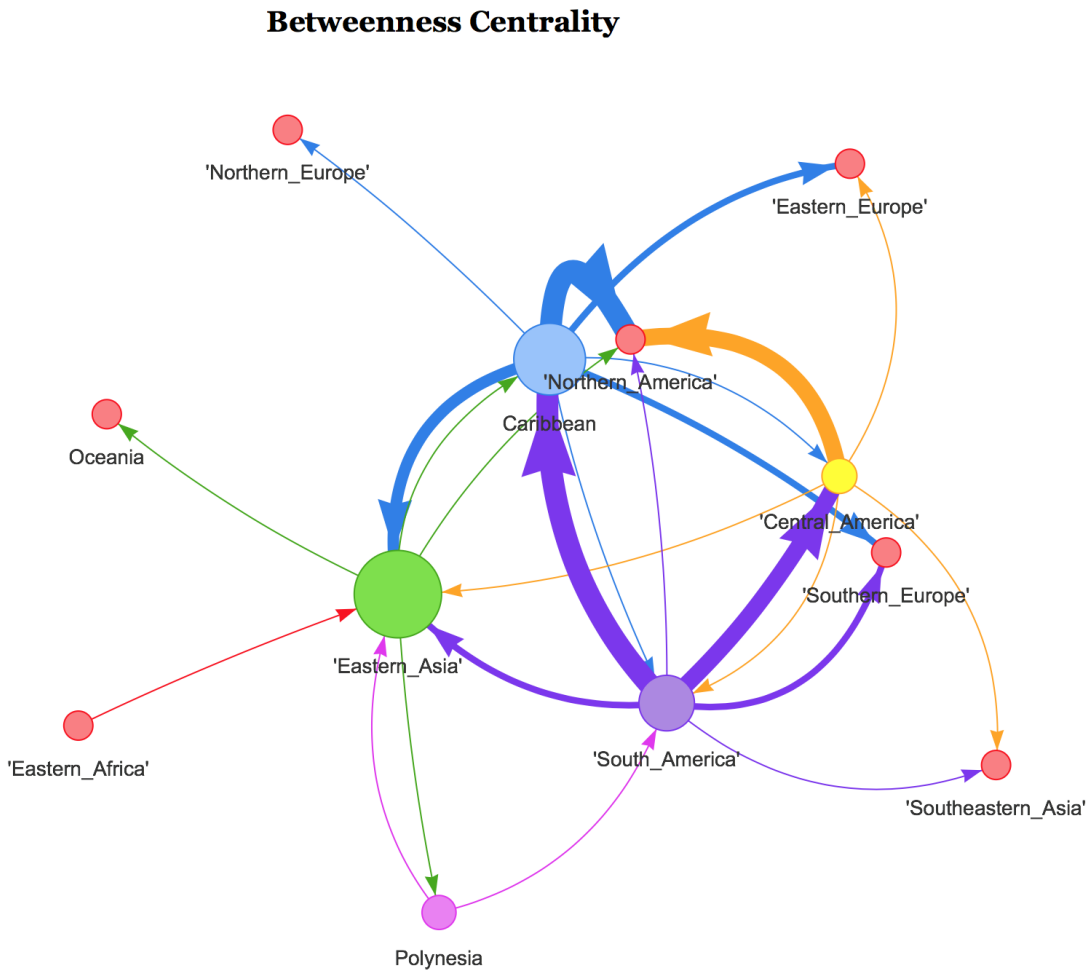

Figure 3.18: Transmission Network of Zika virus Asia Pacific American lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

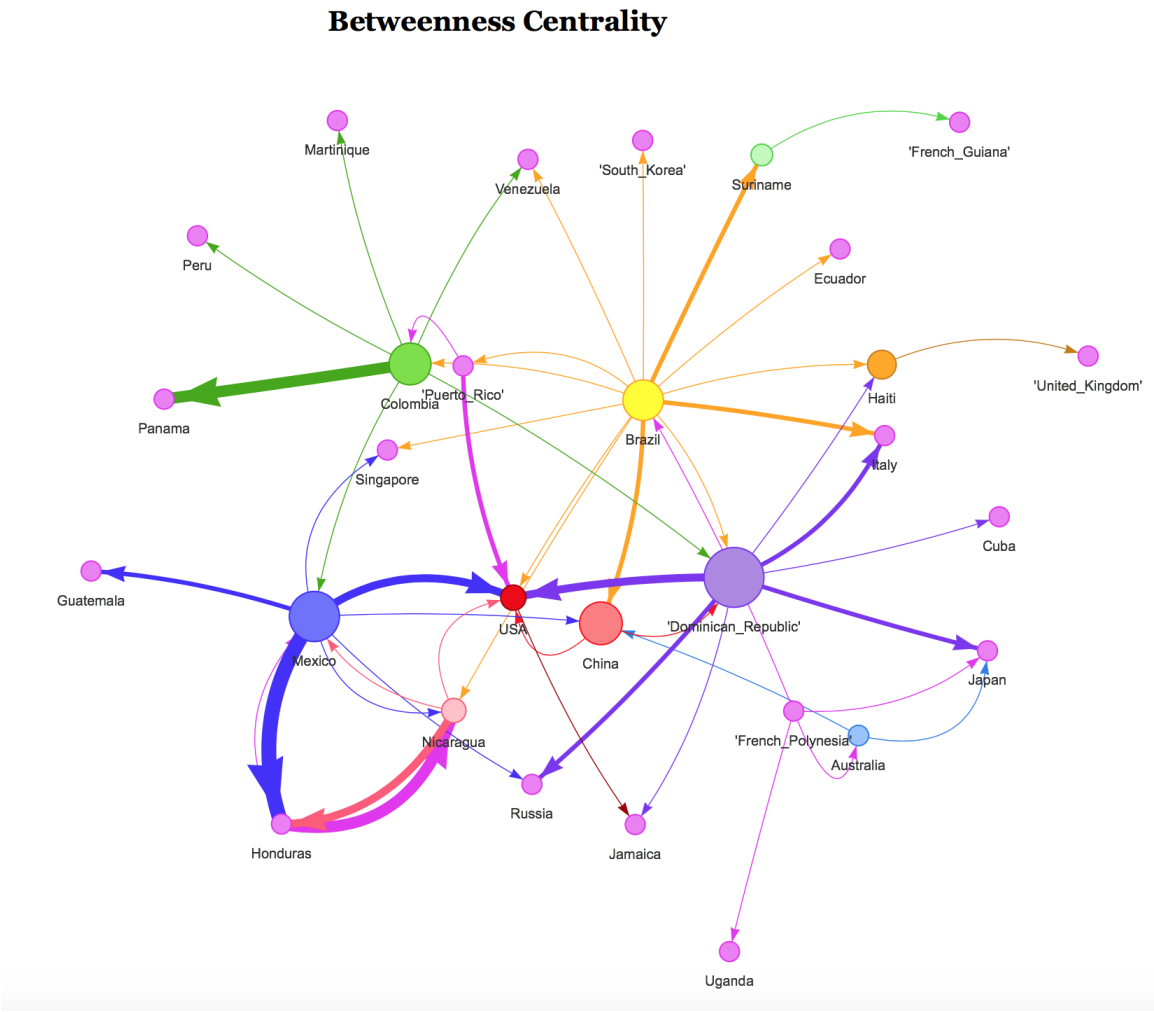

Figure 3.19: Transmission Network of Zika virus Asia Pacific American lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

ZIKV African and Asian dataset was formed with 166 genomic sequences, with 116 sequences originated from Singapore. On the transmission network, the main hub is Southeastern Asia (Figure 3.20). Looking in depth, the main hub in the Eastern Hemisphere is Micronesia, followed by Thailand and Malaysia, while in Africa the relationship between the old African sequences points Uganda as the main hub (Figure 3.21). The presence of 116 sequences from Singapore but no role within the transmission graph can be explained as there is little variation observed among the

strains, isolated from a single outbreak (Figure 2.8).

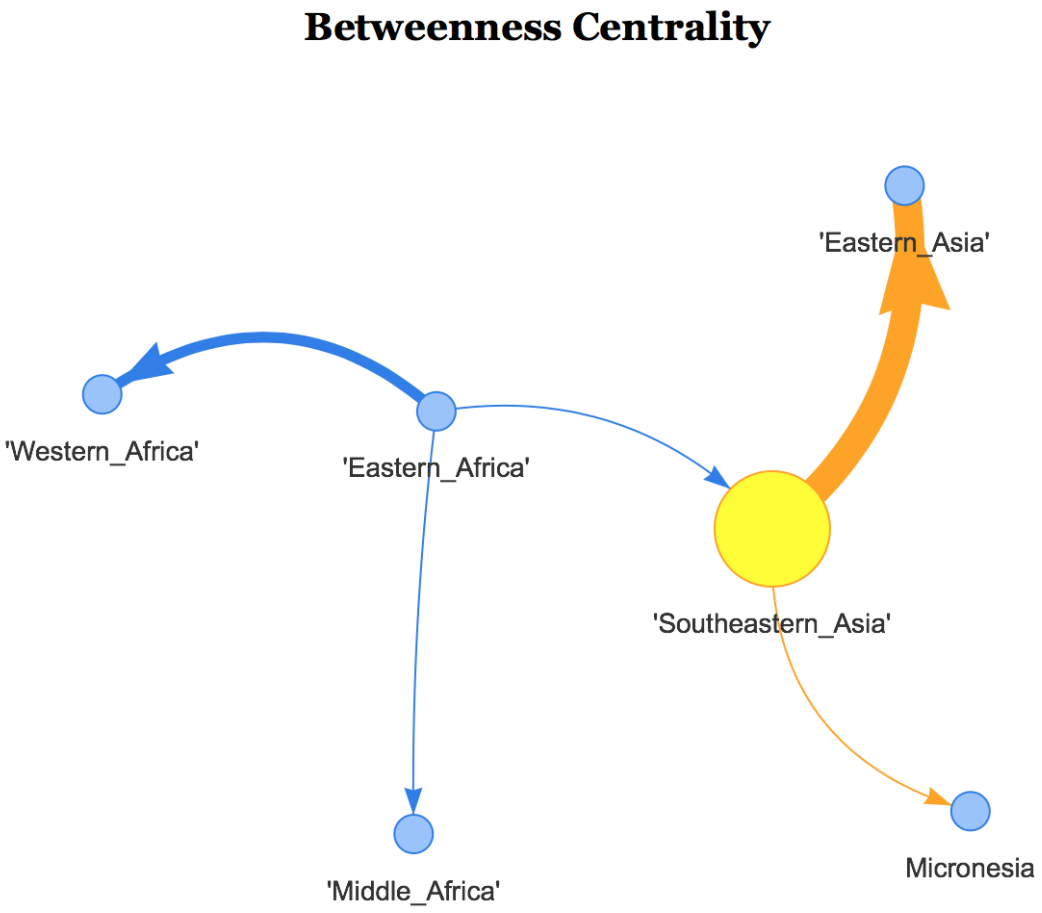

Figure 3.20: Transmission Network of Zika virus African strains and Asian lineage. Nodes represent geographical regions according to UN Geoscheme and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

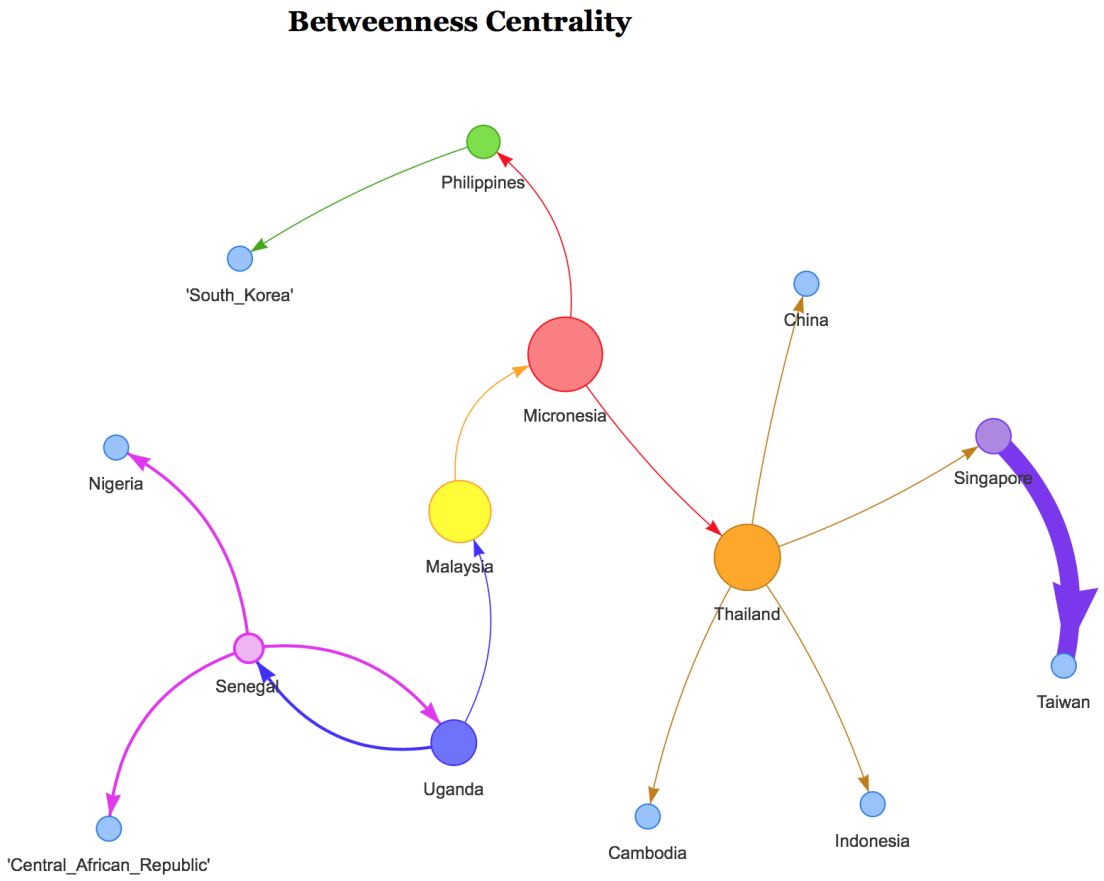

Figure 3.21: Transmission Network of Zika virus African strains and Asian lineage. Nodes represent countries and are scaled by betweenness centrality metric (larger = more betweenness). Width of ties is scaled by number of shifts between nodes (larger = more frequent), colors represent assigned metric values (same color = same value).

Table 3.4: Summary of lineage specific transmission networks results.

Virus	Georegion lineage main hub(s)	Country lineage main hub(s)
Chikungunya Asian Urban	Caribbean; Polynesia	Caribbean Islands; Nicaragua and USA
Chikungunya Indian Ocean Lineage	Southern Asia; Southeastern Asia	Malaysia; Thailand and India
Chikungunya South American Lineage	Eastern Africa; Southern Asia	India; Angola
Chikungunya West African Lineage	West Africa	Senegal
Zika Asia Pacific American Lineage	Eastern Asia; Caribbean; South America	Dominican Republic; Mexico; Colombia; Brazil and China
Zika African Strains Asian Lineage	Southeastern Asia	Micronesia; Malaysia and Thailand

3.3.3 Historical Transmission Network Comparisons

In order to investigate the possible correlation between the different transmission networks and also to evaluate if there is a relationship between the different metrics applied in this study given that they are for the first time being tested in this scenario to infer epidemiological meaning to a network, I compared the historical transmission networks in two distinct ways. One, how the transmission networks correlate using distinct centrality metrics (Table 3.5, 3.6, 3.7 and 3.8), and two, how the distinct metrics correlate to each other given the current transmission networks available (Table 3.22). For the purpose of comparison, correlation coefficient >0.7 was considered strong, between 0.3-0.7 moderate and <0.3 weak.

Betweenness, closeness and SHR had a strong correlation between the different DENV serotypes (Tables 3.5,3.6 and 3.8). CHIKV had the lowest correlations between all viruses on the three metrics, followed by YFV. ZIKV and CHIKV had a moderate

positive betweenness centrality correlation, which could indicate shared hubs, but this, of course, requires further evidence.

Degree had strong correlation overall, except for YFV (Table 3.7). As Degree expresses the overall number of shifts in and out of the locations, it seems that there is on average a strong positive correlation between the rank of geographic regions and CHIKV, DENV-1, DENV-2, DENV-3, DENV-4 and ZIKV. This only means though that in most of the networks, there is a tendency to the same regions to have an elevated number of changes within the network.

Table 3.5: Betweenness Centrality correlation comparison.

Virus	CHIKV	DENV-1	DENV-2	DENV-3	DENV-4	YFV	ZIKV
CHIKV	1.00	-0.19	0.20	0.07	0.10	0.03	0.37
DENV-1	-0.19	1.00	0.53	0.70	0.56	0.15	0.50
DENV-2	0.20	0.53	1.00	0.66	0.77	0.12	0.72
DENV-3	0.20	0.70	0.66	1.00	0.67	0.34	0.58
DENV-4	0.10	0.56	0.77	0.67	1.00	0.25	0.56
YFV	0.03	0.15	0.12	0.34	0.25	1.00	0.06
ZIKV	0.37	0.50	0.72	0.58	0.56	0.06	1.00

Table 3.6: Closeness Centrality correlation comparison.

Virus	CHIKV	DENV-1	DENV-2	DENV-3	DENV-4	YFV	ZIKV
CHIKV	1.00	-0.20	0.00	-0.13	-0.16	-0.03	0.15
DENV-1	-0.20	1.00	0.62	0.82	0.65	0.30	0.26
DENV-2	0.00	0.62	1.00	0.86	0.74	0.05	0.45
DENV-3	-0.13	0.82	0.86	1.00	0.78	0.12	0.39
DENV-4	-0.16	0.65	0.74	0.78	1.00	0.26	0.30
YFV	-0.03	0.30	0.05	0.12	0.26	1.00	0.17
ZIKV	0.15	0.26	0.45	0.39	0.30	0.17	1.00

Table 3.7: Degree Centrality correlation comparison.

Virus	CHIKV	DENV-1	DENV-2	DENV-3	DENV-4	YFV	ZIKV
CHIKV	1.00	0.86	0.65	0.80	0.70	0.18	0.55
DENV-1	0.86	1.00	0.69	0.89	0.78	0.09	0.52
DENV-2	0.65	0.69	1.00	0.87	0.82	0.15	0.78
DENV-3	0.80	0.89	0.87	1.00	0.90	0.09	0.70
DENV-4	0.70	0.78	0.82	0.90	1.00	0.11	0.52
YFV	0.18	0.09	0.15	0.09	0.11	1.00	0.32
ZIKV	0.55	0.52	0.78	0.70	0.52	0.32	1.00

Table 3.8: SHR correlation comparison.

Virus	CHIKV	DENV-1	DENV-2	DENV-3	DENV-4	YFV	ZIKV
CHIKV	1.00	-0.26	-0.10	-0.20	-0.27	0.16	0.27
DENV-1	-0.26	1.00	0.67	0.75	0.65	0.15	0.50
DENV-2	-0.10	0.67	1.00	0.85	0.69	0.17	0.63
DENV-3	-0.20	0.75	0.85	1.00	0.77	-0.14	0.49
DENV-4	-0.27	0.65	0.69	0.77	1.00	-0.14	0.31
YFV	0.16	0.15	0.17	-0.14	-0.14	1.00	0.39
ZIKV	0.27	0.50	0.63	0.49	0.31	0.39	1.00

By taking a look on the boxplot of the correlation coefficient between different metrics using the historical transmission network datasets (Figure 3.22), it can be observed that there is a relatively strong positive correlation for all the metrics studied. Curiously, DENV-4 and YFV had low negative correlation values for Degree x Indegree (DxI), which enlarged the size of variation of correlation coefficient within DxI.

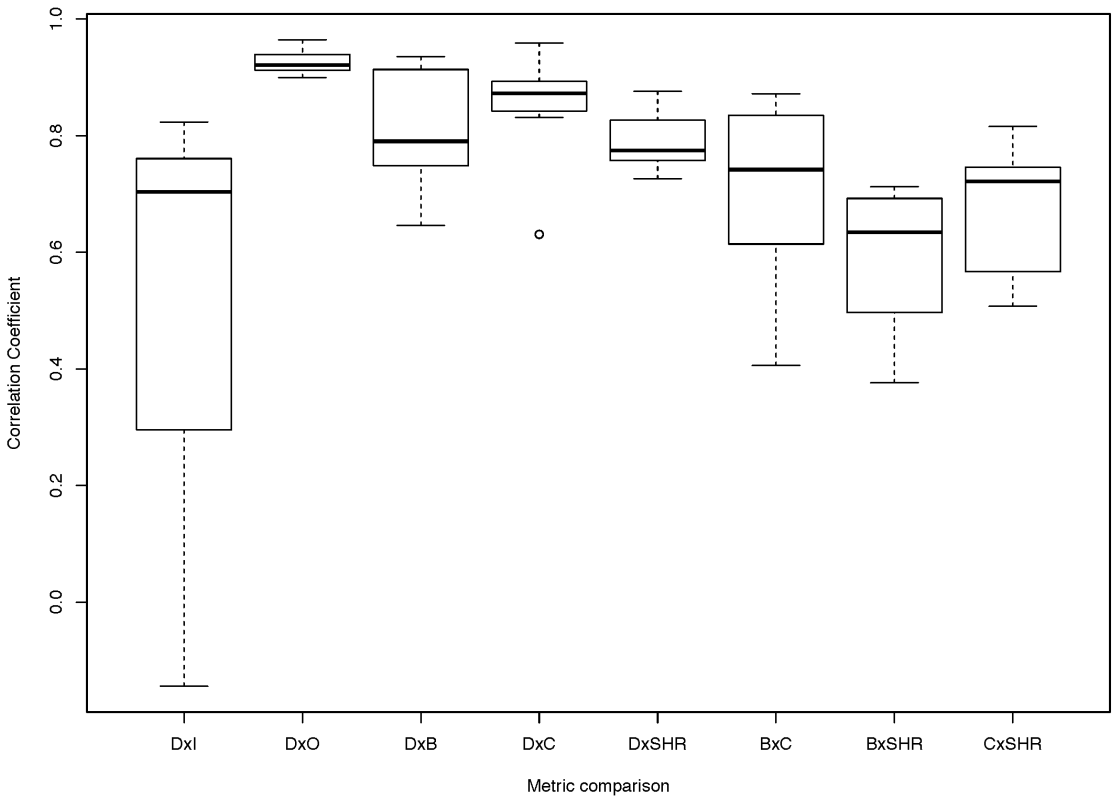

Figure 3.22: Boxplot of the correlation coefficient between different metrics given the current datasets. B = Betweenness centrality / C = Closeness centrality / D = Degree centrality / I = Indegree / O = Outdegree / SHR = Source Hub Ratio.

3.4 Discussion

Pathogen transmission networks requires a good initial dataset in order for the network to have epidemiological meaning. Building transmission networks from phylogenetic trees has been done multiple times in the past using different strategies and

its efficacy been argued [131]. Famulare et al. [132] conclude in a study that the spatial history reconstruction utilizing phylogenetics is limited by under-sampling [132]. Nonetheless, in the past years better sampling and sequencing strategies have started to provide very densely sampled datasets [133].

In order to understand how diseases spread, traditional epidemiological models such as susceptible-infectious-recovered (SIR) and susceptible-infectious-susceptible (SIS) have been used to describe change in terms of case counts in each of the different infectious states of the hosts. The flow of numbers of cases in each state describes the pace of the epidemic, termed basic reproductive number [134].

Other forms of epidemiological studies focus on the interactions between the hosts in order to build transmission networks to describe the spread of the disease [135]. Building a transmission network out of contact tracing data depends on multiple factors, such as the quality of the data collection, the difficulty of linking individuals within the network, and the difficulty of determining polarity of host to host transmission events [136]. Pulse-field gel electrophoresis and variants of DNA fingerprinting been used by epidemiologists in order to to identify different strains. Recently, whole genome sequencing (WGS) has supplanted these technologys because WGS affords higher resolution of pathogen lineages [137, 138, 139].

The results of the current work is not to focus on contact tracing but rather pathogen genomes and metadata with WGS or partial genomes. I have shown that by evaluating the relationship between the pathogen based on genetics, transmission events can be inferred. In other words, polarized transmission events can be traced among viral lineages which allows us to build a directed network and reconstruct the movement of the pathogen over hosts and or geography.

Alternative approaches such as data mining on social media in order to supplement existing reviews or analysis have been performed but with limited success [140, 141, 142]. Some of the problems stem from unstandardized sources of data, mul-

tiple sources provide summarized information that inhibits the analysis to be more informative [143]. By utilizing only the public data currently available for four neglected tropical diseases, CHIKV, DENV, ZIKV and YFV, I was able to reconstruct the epidemiological history of these viruses as well as point important geographic regions by applying network metrics.

CHIKV historical transmission behavior shows that there is a major corridor in Southeastern and Southern Asia responsible for the massive spread of the disease. When looking at the individual outbreaks we see that this behavior is driven by a specific lineage, the Indian Ocean Lineage. The Asian Urban Lineage that reached the Americas shifted the hub centrality to the Americas. This raises the importance of the study of individual lineages rather than simply looking at large historical transmission networks. While this is true to CHIKV, the transmission network of DENV, across serotype, had a balance of major hubs between Southeastern, Southern Asia and the Americas. These results correlate with the current geographic distribution of DENV and epidemiological data [144].

Results for YFV are the example of a virus that although been circulating in the Americas for hundreds of years, has been neglected under poor surveillance when compared to other viruses such as DENV and CHIKV. Only this year, in the light of new technologies and partnerships, have researchers sequenced and investigated an epidemic in detail in Brazil [145]. However, the lack of background data did not allow a strong correlation between the epidemiological history and my results as YFV is endemic in both South America and Africa [146].

As mentioned previously, ZIKV has a poor history in Africa. Sequences found in Africa were isolated from multiple vectors and number of sequences available is small. ZIKV also have had a recent and short history in Asia as a new lineage. Although some samples of historical Asian lineages are available for ZIKV, the lack of sampling complicates the reconstruction of outbreaks of this virus. Most of what can be seen

in the transmission network is from sequences with low genetic distance due to the past epidemic (Appendix Figure A.2).

The recent outbreak of ZIKV has shown that South America, more specifically Brazil served as a hub and source for the spread of the epidemic. ZIKV lineages in Brazil is connected to the Central America and Caribbean lineages of ZIKV. Brazil region which had largest betweenness centrality for ZIKV thus seem to be the best hub for this epidemic.

The hub in Asia being China for the ZIKV Asia Pacific American lineage raised the question of what was the reason for China to be a hub as there were no reports of local transmission of ZIKV. Traveler cases been described in the literature for China as well as other locations such as Russia and the United Kingdom but no secondary transmission was identified. Thus, it appears that the importance of China within the network is due to the similarity of the sequence of the original place, were the traveler acquired the disease, to sequences to other locations and should be disregarded.

There is a limitation on the analyses based on the data quality and availability given that there were no recurrent and standardized viral sampling in all regions of the world which had been affected by these diseases. I can only describe the relationships between the sampled data and infer they are connected, even though there could be an intermediary node missing due to under sampling.

3.4.1 Comparing Networks

Multiple works on network comparison focused on graph matching, calculating the distance of the actual shape of the network and not the metrics extracted from them [147]. Rather of comparing how distant the networks are in terms of shape, specially given that they do not share the same connected nodes, I compared the rank assigned to their nodes using the different metrics described on Table 3.2.

Simple comparison metrics such as the Jaccard coefficient have been used in previous studies to compare centrality metrics within networks [148]. Given that nodes

within the transmission network in this case are geographic regions/countries, the assumption that nodes with no shifts do exist is valid and thus I can simply assume that nodes not present scored zero in all the metrics. This scoring indicates that there was no shift of state nor presence of the disease in that geographic location. By doing this, I was allowed to perform a simple Pearson's correlation in order to evaluate the strength of the relationship between the ranks between the networks as all nodes are present in the networks, but are not necessarily connected. Other metrics to compare the networks may be used. In this work, I found that due to geography the application of Pearson correlation has the best epidemiological meaning.

When comparing networks, it only made sense to do it between different viruses, given that the known lineages had not similar geographic spread. Different metrics had strong positive correlation. Interestingly, DENV serotypes had the strongest correlations overall, which makes sense according to their epidemiology. It is not uncommon to see multiple if not all four DENV serotypes co-circulating in endemic areas. For example, in Malaysia, at least three serotypes, DENV-2,DENV-3 and DENV-4 co-circulate [149]. In Brazil co-circulation of all four serotypes has been reported [150]. A report from 2017 also mentions the co-circulation of all four DENV serotypes in India and calls the attention for better molecular monitoring of circulating serotypes [151].

YF and CHIKV had the lowest correlation with all the other viruses in all different metrics. ZIKV was the third one. CHIKV is a different virus, just included as it is mainly transmitted by the same vector, *Aedes aegypti*. This indicates that although they now coexist in multiple regions in the world, their epidemiological history distance them at this moment. Given that CHIKV and ZIKV are recent introductions in the Americas, the behavior of these viruses within the network should be continuing reevaluated as the virus evolves and more genetic information is added into the databases.

3.4.2 Comparing Metrics

The fact that all viruses had a strong positive correlation with Degree Centrality but not necessarily with all the other metrics raises the question of why and how valid Degree could be within a given network as a good metric to identify important nodes within the network. It seems that by taking into account just the number of shifts originating and arriving to the nodes, Degree Centrality could create an inference bias to map locations where better surveillance has been put in place. Even though the technique is robust as it cares about shifts between metadata states within the network, ignoring multiple sequences with same metadata unless there is a shift, heavier surveillance on individual locations could influence the network.

Also, Degree includes indegree and outdegree, which makes the metric flag nodes within the network based on how many shifts you have coming in and out of the node, ignoring their directionality, which makes it difficult to distinguish important nodes that could be source versus a dead end for the virus transmission. Assuming good standardized surveillance across borders, utilizing indegree and outdegree separately would make this metric optimal. Another metric that takes into account these shifts, Source Hub Ratio (SHR), seems to perform better by not giving a score based on shifts, but on a rate on a range from 0 to 1, which also distinguishes nodes within the networks as of their role. The variable correlation between degree centrality and indegree for each dataset, which is distinct from outdegree was observed in the opposite way in a previous study and was justified as a specific behavior of the dataset utilized to calculate correlation [152].

Betweenness centrality works the best of all metrics evaluated as it looks for hubs within the network assuming the shape of the network and the flux between the nodes. Although it was not the main focus in this work, it seems that Betweenness centrality is affected when there is a low number of sequences, thus, shifts within the network. Betweenness centrality still flags the most important nodes within the network given

the shortest paths in the network, but not necessarily relates back to epidemiology. A novel metric that merges the concept of SHR and Betweenness could be of great value for network analysis.

Closeness centrality was an interesting idea, but it tends to increase the size of too many nodes and it does not necessarily give an input in terms of importance of the node. This happens because a node can be very close to a relatively subset within the network, or moderately close to every node in a larger network and still receive the same closeness score. Degree centrality is good to identify nodes that originate multiple events, but it inflates the nodes based on the number of transmission events (shifts of metadata), possibly introducing a bias on the evaluation of nodes based on sampling. Although this could be true for other metrics, degree gets affected the most as it ranks based on the number of in and out degrees and those are not normalized within the network.

SHR is a novel addition as it takes the idea from degree centrality and transforms it into a rate that makes it easy to see the role of a node within the network. The downside is that since SHR does not account for centrality, you have to use another metric to classify by importance within the network which node should be investigated. SHR values in a small network or applied to scale network nodes does work well to show the function of a node where the size of the node can be distinguished by simply being looked at it, larger complex networks makes the differences between nodes difficult to visualize and centrality is required to filter out nodes with few connections.

CHAPTER 4: CONCLUSION

Current phylogeny of CHIKV is misguided due to nomenclature. More attention should be given to African strains and more studies should be done in order to investigate CHIKV dynamics in Africa. DENV datasets reflect current epidemiology and can be used to identify geographic regions where public health efforts may be intensified in order to reduce transmission events.

ZIKV had a rapid spread and outbreak. ZIKV apparently has a unique mechanism that causes fetal demise utilizing 3' UTR MBEs. It remains to be seen if ZIKV, now endemic in the Americas, will cause recurring outbreaks like CHIKV, DENV and YFV do and if congenital defects will continue to be present. ZIKV in Asia remains to be better understood in terms of spread and presence and if the Asian clade is in current expansion or not.

YFV requires more attention and genomic sequencing. The low background data available for the YFV compared to all other viruses does not allow a fair comparison with other outbreaks, nor a throughout study of the history of epidemics and major hubs.

A question that remains to be answered out of all outbreaks is what triggered the movement of the diseases from one endemic region to another in a specific time? Are outbreaks isolated events or are they triggered by something that could happen over and over again? Who was patient "zero" for that specific outbreak? Theories vary from major sports events causing a mass movement of people from different countries to the host country, bringing diseases not previously seen to a immunologically naïve population, to individual travelers over time being bitten by mosquitoes and that growing exponentially to the point of causing outbreaks. These question remains to

be answered for a vast amount of diseases, specially those labeled "Neglected Tropical Diseases". Better monitoring associated with phylogenetics techniques employed in this work should help to answer it.

The tools built and utilized during this study provide the grounds for a better surveillance system to be put in place. As new and cheaper technologies arrive, more genomic data will allow better and more fine grained studies for arboviruses as well as other Neglected Tropical Diseases. The results on this work open multiple questions regarding the role of specific nodes in different transmission networks and should be investigated further and in depth with additional clinical data.

4.0.1 Significance and Future Work

4.0.1.1 Phylogenetics and genomics

Phylogenetic studies on the evolution of new pathogen outbreaks as new data becomes available has become highly significant to the field of Virology and Epidemiology. Understanding the evolutionary relationships between different strains provide new insights on outbreak behavior which ultimately lead to guidance to health workers in the field on how to proceed with interventions.

I built a pipeline utilizing existing tools and CHIKV and ZIKV as study models to investigate genomic changes over time and space which can now be applied to different pathogens. Future work entails on developing a software which integrates these different tools and data manipulation in a single place.

4.0.1.2 Transmission Networks

The study of pathogen transmission networks is important for understanding the spread of diseases and to identify specific hubs and locations where interventions can be made to halt epidemics from continuing to spread. The ability to correlate the networks utilizing as starting point not the structure of the network but metrics calculated from the networks allows us to identify patterns among different viruses

and between separate outbreaks.

Transmission networks scaled by betweenness centrality point to main hubs within networks. An association between SHR and betweenness centrality may be necessary to increase robustness of analysis to confirm important nodes within networks as it associates the rate of movement out of the node with the node acting as the optimal intermediary node between other nodes within network. Degree centrality seems to be too raw to give informative epidemiological insights, as it basically relies on the count of number of events in and out of a node. Closeness centrality also did not seem like a good epidemiological indicator as it gives similar scores to distinct node behaviors within the network.

Future work entails the development of a web-based server to make StrainHub publicly available and the implementation and test of different and novel metrics as well as the graphic interface and other visualization tools to improve network visualization and characterization.

REFERENCES

[1] N. R. Zearfoss, L. M. Deveau, C. C. Clingman, E. Schmidt, E. S. Johnson, F. Massi, and S. P. Ryder, "A conserved three-nucleotide core motif defines Musashi RNA binding specificity," *J Biolo Chem*, vol. 289, no. 51, pp. 35530–35541, 2014.

[2] A. de Bernardi Schneider, R. W. Malone, J.-T. Guo, J. Homan, G. Linchangco, Z. L. Witter, D. Vinesett, L. Damodaran, and D. A. Janies, "Molecular evolution of Zika virus as it crossed the Pacific to the Americas," *Cladistics*, vol. 33, no. 1, pp. 1–20, 2017.

[3] P. J. Hotez and K. O. Murray, "Dengue, West Nile virus, Chikungunya, Zika and now Mayaro?," *PLoS neglected tropical diseases*, vol. 11, no. 8, p. e0005462, 2017.

[4] J. Tognarelli, S. Ulloa, E. Villagra, J. Lagos, C. Aguayo, R. Fasce, B. Parra, J. Mora, N. Becerra, N. Lagos, *et al.*, "A report on the outbreak of Zika virus on Easter Island, South Pacific, 2014," *Archives of virology*, vol. 161, no. 3, pp. 665–668, 2016.

[5] W. Lumsden, "An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953 II. General description and epidemiology," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 49, no. 1, pp. 33–57, 1955.

[6] I. Leparc-Goffart, A. Nougairede, S. Cassadou, C. Prat, and X. De Lamballerie, "Chikungunya in the Americas," *The Lancet*, vol. 383, no. 9916, p. 514, 2014.

[7] G. Dick, S. Kitchen, and A. Haddow, "Zika virus (I). isolations and serological specificity," *Transactions of the royal society of tropical medicine and hygiene*, vol. 46, no. 5, pp. 509–520, 1952.

[8] "Timeline of the emergence of Zika virus in the Americas," Apr 2016. www.paho.org.

[9] M. C. Robinson, "An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 49, no. 1, pp. 28–32, 1955.

[10] X.-F. Li, T. Jiang, Y.-Q. Deng, H. Zhao, X.-D. Yu, Q. Ye, H.-J. Wang, S.-Y. Zhu, F.-C. Zhang, E.-D. Qin, *et al.*, "Complete genome sequence of a Chikungunya virus isolated in guangdong, china," *Journal of virology*, vol. 86, no. 16, pp. 8904–8905, 2012.

[11] G. Borgherini, P. Poubeau, F. Staikowsky, M. Lory, N. L. Moullec, J. P. Becquart, C. Wengling, A. Michault, and F. Paganin, "Outbreak of Chikungunya

on Reunion Island: early clinical and laboratory features in 157 adult patients," *Clinical infectious diseases*, vol. 44, no. 11, pp. 1401–1407, 2007.

[12] R. Pulmanausahakul, S. Roytrakul, P. Auewarakul, and D. R. Smith, "Chikungunya in Southeast Asia: understanding the emergence and finding solutions," *International Journal of Infectious Diseases*, vol. 15, no. 10, pp. e671–e676, 2011.

[13] A. Suhrbier and M. La Linn, "Clinical and pathologic aspects of arthritis due to Ross River virus and other alphaviruses," *Current opinion in rheumatology*, vol. 16, no. 4, pp. 374–379, 2004.

[14] S. C. Weaver and M. Lecuit, "Chikungunya virus and the global spread of a mosquito-borne disease," *New England Journal of Medicine*, vol. 372, no. 13, pp. 1231–1239, 2015.

[15] R. Ross, "The Newala epidemic: III. the virus: isolation, pathogenic properties and relationship to the epidemic," *Epidemiology & Infection*, vol. 54, no. 2, pp. 177–191, 1956.

[16] P. Mason and A. Haddow, "An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952–1953: An additional note on Chikungunya virus isolations and serum antibodies," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 51, no. 3, pp. 238–240, 1957.

[17] R. M. Langsjoen, S. L. Haller, C. J. Roy, H. Vinet-Oliphant, N. A. Bergren, J. H. Erasmus, J. A. Livengood, T. D. Powell, S. C. Weaver, and S. L. Rossi, "Chikungunya virus strains show lineage-specific variations in virulence and cross-protective ability in murine and nonhuman primate models," *mBio*, vol. 9, no. 2, pp. e02449–17, 2018.

[18] H. Zeller, W. Van Bortel, and B. Sudre, "Chikungunya: its history in Africa and Asia and its spread to new regions in 2013–2014," *The Journal of infectious diseases*, vol. 214, no. suppl_5, pp. S436–S440, 2016.

[19] D. E. Carey, "Chikungunya and Dengue: a case of mistaken identity?," *Journal of the history of medicine and allied sciences*, vol. 26, no. 3, pp. 243–262, 1971.

[20] A. M. Powers, A. C. Brault, R. B. Tesh, and S. C. Weaver, "Re-emergence of Chikungunya and O'nyong-nyong viruses: evidence for distinct geographical lineages and distant evolutionary relationships," *Journal of General Virology*, vol. 81, no. 2, pp. 471–479, 2000.

[21] S. S. Cherian, A. M. Walimbe, S. M. Jadhav, S. S. Gandhe, S. L. Hundekar, A. C. Mishra, and V. A. Arankalle, "Evolutionary rates and timescale comparison of Chikungunya viruses inferred from the whole genome/e1 gene with special reference to the 2005–07 outbreak in the Indian subcontinent," *Infection, Genetics and Evolution*, vol. 9, no. 1, pp. 16–23, 2009.

[22] P. Gérardin, V. Guernier, J. Perrau, A. Fianu, K. Le Roux, P. Grivard, A. Michault, X. De Lamballerie, A. Flahault, and F. Favier, "Estimating Chikungunya prevalence in La Reunion Island outbreak by serosurveys: two methods for two critical times of the epidemic," *BMC infectious diseases*, vol. 8, no. 1, p. 99, 2008.

[23] L. Josseran, C. Paquet, A. Zehgnoun, N. Caillere, A. T. Le, J.-L. Solet, and M. Ledrans, "Chikungunya disease outbreak, Reunion Island.," *Emerging infectious diseases*, vol. 12, no. 12, pp. 1994–1995, 2006.

[24] K. A. Tsetsarkin, D. L. Vanlandingham, C. E. McGee, and S. Higgs, "A single mutation in Chikungunya virus affects vector specificity and epidemic potential," *PLoS pathogens*, vol. 3, no. 12, p. e201, 2007.

[25] M. R. T. Nunes, N. R. Faria, J. M. de Vasconcelos, N. Golding, M. U. Kraemer, L. F. de Oliveira, R. d. S. da Silva Azevedo, D. E. A. da Silva, E. V. P. da Silva, S. P. da Silva, *et al.*, "Emergence and potential for spread of Chikungunya virus in Brazil," *BMC medicine*, vol. 13, no. 1, p. 102, 2015.

[26] A. D. Haddow, A. J. Schuh, C. Y. Yasuda, M. R. Kasper, V. Heang, R. Huy, H. Guzman, R. B. Tesh, and S. C. Weaver, "Genetic characterization of Zika virus strains: geographic expansion of the Asian lineage," *PLoS neglected tropical diseases*, vol. 6, no. 2, p. e1477, 2012.

[27] J. Olson, T. Ksiazek, *et al.*, "Zika virus, a cause of fever in Central Java, Indonesia," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 75, no. 3, pp. 389–393, 1981.

[28] N. Marchette, R. Garcia, and A. Rudnick, "Isolation of Zika virus from Aedes aegypti mosquitoes in Malaysia," *The American journal of tropical medicine and hygiene*, vol. 18, no. 3, pp. 411–415, 1969.

[29] B. Parra, J. Lizarazo, J. A. Jiménez-Arango, A. F. Zea-Vera, G. González-Manrique, J. Vargas, J. A. Angarita, G. Zuñiga, R. Lopez-Gonzalez, C. L. Beltran, *et al.*, "Guillain–Barré syndrome associated with Zika virus infection in Colombia," *New England Journal of Medicine*, vol. 375, no. 16, pp. 1513–1523, 2016.

[30] S. Cauchemez, M. Besnard, P. Bompard, T. Dub, P. Guillemette-Artur, D. Eyrolle-Guignot, H. Salje, M. D. Van Kerkhove, V. Abadie, C. Garel, *et al.*, "Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study," *The Lancet*, vol. 387, no. 10033, pp. 2125–2132, 2016.

[31] T. Jaenisch, K. D. Rosenberger, C. Brito, O. Brady, P. Brasil, and E. T. Marques, "Risk of microcephaly after Zika virus infection in Brazil, 2015 to 2016," *Bulletin of the World Health Organization*, vol. 95, no. 3, p. 191, 2017.

[32] Z. A. Klase, S. Khakhina, A. D. B. Schneider, M. V. Callahan, J. Glasspool-Malone, and R. Malone, "Zika fetal neuropathogenesis: etiology of a viral syndrome," *PLoS neglected tropical diseases*, vol. 10, no. 8, p. e0004877, 2016.

[33] E. Oehler, L. Watrin, P. Larre, I. Leparc-Goffart, S. Lastere, F. Valour, L. Baudouin, H. Mallet, D. Musso, and F. Ghawche, "Zika virus infection complicated by Guillain-Barre syndrome–case report, French Polynesia, December 2013," *Eurosurveillance*, vol. 19, no. 9, p. 20720, 2014.

[34] "Rapid risk assessment Zika virus disease epidemic," Oct 2016.

[35] G. Calvet, R. S. Aguiar, A. S. Melo, S. A. Sampaio, I. De Filippis, A. Fabri, E. S. Araujo, P. C. de Sequeira, M. C. de Mendonça, L. de Oliveira, *et al.*, "Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study," *The Lancet infectious diseases*, vol. 16, no. 6, pp. 653–660, 2016.

[36] J. Mlakar, M. Korva, N. Tul, M. Popović, M. Poljšak-Prijatelj, J. Mraz, M. Kolenc, K. Resman Rus, T. Vesnaver Vipotnik, V. Fabjan Vodušek, *et al.*, "Zika virus associated with microcephaly," *New England Journal of Medicine*, vol. 374, no. 10, pp. 951–958, 2016.

[37] L. M. Paul, E. R. Carlin, M. M. Jenkins, A. L. Tan, C. M. Barcellona, C. O. Nicholson, S. F. Michael, and S. Isern, "Dengue virus antibodies enhance Zika virus infection," *Clinical & translational immunology*, vol. 5, no. 12, 2016.

[38] B. Kirya, L. Mukwaya, and S. Sempala, "A Yellow fever epizootic in Ziika Forest, Uganda, during 1972: Part 1: Virus isolation and sentinel monkeys," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 71, no. 3, pp. 254–260, 1977.

[39] K. Smithburn, R. Taylor, F. Rizk, and A. Kader, "Immunity to certain arthropod-borne viruses among indigenous residents of egypt," *The American Journal of Tropical Medicine and Hygiene*, vol. 3, no. 1, pp. 9–18, 1954.

[40] K. Smithburn, J. Kerr, and P. Gatne, "Neutralizing antibodies against certain viruses in the sera of residents of India," *The Journal of Immunology*, vol. 72, no. 4, pp. 248–257, 1954.

[41] K. Smithburn *et al.*, "Neutralizing antibodies against arthropod-borne viruses in the sera of long-time residents of Malaya and Borneo.," *American journal of hygiene*, vol. 59, no. 2, pp. 157–63, 1954.

[42] W. L. Pond, "Arthropod-borne virus antibodies in sera from residents of South-East Asia," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 57, no. 5, pp. 364–371, 1963.

[43] W. M. Hammon, W. Schrack Jr, and G. Sather, "Serological survey for arthropod-borne virus infections in the Philippines," *The American journal of tropical medicine and hygiene*, vol. 7, no. 3, pp. 323–328, 1958.

[44] S. C. Weaver, F. Costa, M. A. Garcia-Blanco, A. I. Ko, G. S. Ribeiro, G. Saade, P.-Y. Shi, and N. Vasilakis, "Zika virus: History, emergence, biology, and prospects for control," *Antiviral research*, vol. 130, pp. 69–80, 2016.

[45] O. Faye, C. C. Freire, A. Iamarino, O. Faye, J. V. C. de Oliveira, M. Diallo, P. M. Zanotto, *et al.*, "Molecular evolution of Zika virus during its emergence in the 20th century," *PLoS neglected tropical diseases*, vol. 8, no. 1, p. e2636, 2014.

[46] C. Zanluca, V. C. A. d. Melo, A. L. P. Mosimann, G. I. V. d. Santos, C. N. D. d. Santos, and K. Luz, "First report of autochthonous transmission of Zika virus in Brazil," *Memórias do Instituto Oswaldo Cruz*, vol. 110, no. 4, pp. 569–572, 2015.

[47] R. S. Lanciotti, O. L. Kosoy, J. J. Laven, J. O. Velez, A. J. Lambert, A. J. Johnson, S. M. Stanfield, and M. R. Duffy, "Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007," *Emerging infectious diseases*, vol. 14, no. 8, p. 1232, 2008.

[48] V.-M. Cao-Lormeau, C. Roche, A. Teissier, E. Robin, A.-L. Berry, H.-P. Mallet, A. A. Sall, and D. Musso, "Zika virus, French Polynesia, South Pacific, 2013," *Emerging infectious diseases*, vol. 20, no. 6, p. 1085, 2014.

[49] D. Musso, "Zika virus transmission from French Polynesia to Brazil," *Emerg Infect Dis*, vol. 21, no. 10, p. 1887, 2015.

[50] R. S. Lanciotti, A. J. Lambert, M. Holodniy, S. Saavedra, and L. d. C. C. Signor, "Phylogeny of Zika virus in western hemisphere, 2015," *Emerging infectious diseases*, vol. 22, no. 5, p. 933, 2016.

[51] A. Perkasa, F. Yudhaputri, S. Haryanto, R. F. Hayati, C. N. Ma'roef, U. Antonjaya, B. Yohan, K. S. A. Myint, J. P. Ledermann, R. Rosenberg, *et al.*, "Isolation of Zika virus from febrile patient, indonesia," *Emerging infectious diseases*, vol. 22, no. 5, p. 924, 2016.

[52] J. Lednicky, V. M. B. De Rochars, M. El Badry, J. Loeb, T. Telisma, S. Chavannes, G. Anilis, E. Cella, M. Ciccozzi, M. Rashid, *et al.*, "Zika virus outbreak in Haiti in 2014: molecular and clinical data," *PLoS neglected tropical diseases*, vol. 10, no. 4, p. e0004687, 2016.

[53] N. R. Faria, R. d. S. da Silva Azevedo, M. U. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Thézé, M. B. Bonsall, T. A. Bowden, I. Rissanen, *et al.*, "Zika virus in the Americas: early epidemiological and genetic findings," *Science*, p. aaf5036, 2016.

[54] A. Casadevall and M. J. Imperiale, "Risks and benefits of gain-of-function experiments with pathogens of pandemic potential, such as influenza virus: a call for a science-based discussion," 2014.

[55] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Molecular biology and evolution*, vol. 32, no. 1, pp. 268–274, 2014.

[56] D. P. Martin, B. Murrell, M. Golden, A. Khoosal, and B. Muhire, "RDP4: Detection and analysis of recombination patterns in virus genomes," *Virus evolution*, vol. 1, no. 1, 2015.

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

[58] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, *et al.*, "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data," *Bioinformatics*, vol. 28, no. 12, pp. 1647–1649, 2012.

[59] W. Maddison and D. Maddison, "Mesquite: a modular system for evolutionary analysis. version 3.04. 2015," 2016.

[60] A. Rambaut, "Figtree," 2016. http://tree.bio.ed.ac.uk/software/figtree/.

[61] D. Janies, "Nvector," 2018. https://github.com/supramap/nvector.

[62] D. J. Machado, "YBYRA facilitates comparison of large phylogenetic trees," *BMC bioinformatics*, vol. 16, no. 1, p. 204, 2015.

[63] G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V. A. Gennarino, D. S. Horner, G. Pavesi, E. Picardi, *et al.*, "UTRdb and UTRsite (release 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D75–D80, 2009.

[64] T.-H. Chang, H.-Y. Huang, J. B.-K. Hsu, S.-L. Weng, J.-T. Horng, and H.-D. Huang, "An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs," vol. 14, no. 2, p. S4, 2013.

[65] P. L. Chavali, L. Stojic, L. W. Meredith, N. Joseph, M. S. Nahorski, T. J. Sanford, T. R. Sweeney, B. A. Krishna, M. Hosmillo, A. E. Firth, *et al.*, "Neurodevelopmental protein Musashi 1 interacts with the Zika genome and promotes viral replication," *Science*, p. eaam9243, 2017.

[66] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 26, 2011.

[67] M. T. Wolfinger, J. Fallmann, F. Eggenhofer, and F. Amman, "ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines," *F1000Research*, vol. 4, no. 50, 2015.

[68] M. Jiang, J. Anderson, J. Gillespie, and M. Mayne, "uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts," *BMC Bioinformatics*, vol. 9, no. 1, p. 192, 2008.

[69] D. Janies, F. Habib, B. Alexandrov, A. Hill, and D. Pol, "Evolution of genomes, host shifts and the geographic spread of SARS-CoV and related coronaviruses," *Cladistics*, vol. 24, no. 2, pp. 111–130, 2008.

[70] K. B. Gibney, M. Fischer, H. E. Prince, L. D. Kramer, K. St. George, O. L. Kosoy, J. J. Laven, and J. E. Staples, "Chikungunya fever in the United States: a fifteen year review of cases," *Clinical infectious diseases*, vol. 52, no. 5, pp. e121–e126, 2011.

[71] H. Cordel, I. Quatresous, C. Paquet, and E. Couturier, "Imported cases of Chikungunya in metropolitan france, april 2005-february 2006," *Weekly releases (1997–2007)*, vol. 11, no. 16, p. 2944, 2006.

[72] G. Rezza, L. Nicoletti, R. Angelini, R. Romi, A. Finarelli, M. Panning, P. Cordioli, C. Fortuna, S. Boros, F. Magurano, *et al.*, "Infection with Chikungunya virus in Italy: an outbreak in a temperate region," *The Lancet*, vol. 370, no. 9602, pp. 1840–1846, 2007.

[73] Z. J. M. Ho, H. C. Hapuarachchi, T. Barkham, A. Chow, L. C. Ng, J. M. V. Lee, Y. S. Leo, K. Prem, Y. H. G. Lim, P. F. de Sessions, *et al.*, "Outbreak of Zika virus infection in Singapore: an epidemiological, entomological, virological, and clinical analysis," *The Lancet Infectious Diseases*, vol. 17, no. 8, pp. 813–821, 2017.

[74] J. H.-O. Pettersson, J. Bohlin, M. Dupont-Rouzeyrol, O. B. Brynildsrud, K. Alfsnes, V.-M. Cao-Lormeau, M. W. Gaunt, A. K. Falconar, X. Lamballerie, V. Eldholm, *et al.*, "Re-visiting the evolution, dispersal and epidemiology of Zika virus in Asia," *Emerging microbes & infections*, vol. 7, no. 1, p. 79, 2018.

[75] N. P. Lindsey, J. E. Staples, K. Powell, I. B. Rabe, M. Fischer, A. M. Powers, O. I. Kosoy, E. C. Mossel, J. L. Munoz-Jordan, M. Beltran, *et al.*, "Ability to serologically confirm recent Zika virus infection in areas with varying past incidence of dengue virus infection-United States and territories, 2016," *Journal of clinical microbiology*, pp. JCM–01115, 2017.

[76] W. K. de Oliveira, "Increase in reported prevalence of microcephaly in infants born to women living in areas with confirmed Zika virus transmission during the first trimester of pregnancy-Brazil, 2015," *MMWR. Morbidity and mortality weekly report*, vol. 65, 2016.

[77] J. Lourenco, M. de Lourdes Monteiro, T. Valdez, J. M. Rodrigues, O. Pybus, and N. R. Faria, "Epidemiology of the Zika virus outbreak in the Cabo Verde Islands, West Africa," *PLoS currents*, vol. 10, 2018.

[78] M. A. H. Braks, N. Honório, L. Lounibos, R. Lourenço-de Oliveira, and S. Juliano, "Interspecific competition between two invasive species of container mosquitoes, Aedes aegypti and Aedes albopictus (Diptera: Culicidae), in Brazil," *Annals of the Entomological Society of America*, vol. 97, no. 1, pp. 130–139, 2004.

[79] G. Spiteri, B. Sudre, A. Septfons, J. Beauté, *et al.*, "Surveillance of Zika virus infection in the EU/EEA, June 2015 to January 2017," *Eurosurveillance*, vol. 22, no. 41, 2017.

[80] F. Amraoui and A.-B. Failloux, "Chikungunya: an unexpected emergence in Europe," *Current opinion in virology*, vol. 21, pp. 146–150, 2016.

[81] E. O. Wiley, "The evolutionary species concept reconsidered," *Systematic zoology*, vol. 27, no. 1, pp. 17–26, 1978.

[82] G. G. Simpson, "Principles of animal taxonomy," 1961.

[83] A. T. Peterson, "Defining viral species: making taxonomy useful," *Virology journal*, vol. 11, no. 1, p. 131, 2014.

[84] M. J. Ward, S. J. Lycett, M. L. Kalish, A. Rambaut, and A. J. L. Brown, "Estimating the rate of intersubtype recombination in early HIV-1 group M strains," *Journal of virology*, vol. 87, no. 4, pp. 1967–1973, 2013.

[85] V. J. Morley, M. G. Noval, R. Chen, S. C. Weaver, M. Vignuzzi, K. A. Stapleford, and P. E. Turner, "Chikungunya virus evolution following a large 3' UTR deletion results in host-specific molecular changes in protein-coding regions," *Virus Evolution*, vol. 4, no. 1, p. vey012, 2018.

[86] L. Yuan, X.-Y. Huang, Z.-Y. Liu, F. Zhang, X.-L. Zhu, J.-Y. Yu, X. Ji, Y.-P. Xu, G. Li, C. Li, *et al.*, "A single mutation in the prm protein of Zika virus contributes to fetal microcephaly," *Science*, vol. 358, no. 6365, pp. 933–936, 2017.

[87] N. D. Grubaugh and K. G. Andersen, "Navigating the zika panic," *F1000Research*, vol. 5, 2016.

[88] R. Lorenz, M. T. Wolfinger, A. Tanzer, and I. L. Hofacker, "Predicting RNA secondary structures from sequence and probing data," *Methods*, vol. 103, pp. 86–98, 2016.

[89] J. L. Hyde, R. Chen, D. W. Trobaugh, M. S. Diamond, S. C. Weaver, W. B. Klimstra, and J. Wilusz, "The 5' and 3' ends of Alphavirus RNAs–non-coding is not non-functional," *Virus research*, vol. 206, pp. 99–107, 2015.

[90] W. C. Ng, R. Soto-Acosta, S. S. Bradrick, M. A. Garcia-Blanco, and E. E. Ooi, "The 5' and 3' untranslated regions of the flaviviral genome," *Viruses*, vol. 9, no. 6, p. 137, 2017.

[91] J. M. Sutherland, N. A. Siddall, G. R. Hime, and E. A. McLaughlin, "RNA binding proteins in spermatogenesis: an in depth focus on the Musashi family," *Asian journal of andrology*, vol. 17, no. 4, p. 529, 2015.

[92] S.-i. Sakakibara, T. Imai, K. Hamaguchi, M. Okabe, J. Aruga, K. Nakajima, D. Yasutomi, T. Nagata, Y. Kurihara, S. Uesugi, *et al.*, "Mouse-Musashi-1, a neural RNA-binding protein highly enriched in the mammalian CNS stem cell," *Developmental biology*, vol. 176, no. 2, pp. 230–242, 1996.

[93] A. de Bernardi Schneider and M. T. Wolfinger, "Musashi binding elements in Zika and related Flavivirus 3'UTRs: A comparative study in silico," *bioRxiv*, 2018.

[94] D. J. Platt, A. M. Smith, N. Arora, M. S. Diamond, C. B. Coyne, and J. J. Miner, "Zika virus-related neurotropic flaviviruses infect human placental explants and cause fetal demise in mice," *Sci Transl Med*, vol. 10, no. 426, p. eaao7090, 2018.

[95] J. R. Torres, L. H. Falleiros-Arlant, L. Dueñas, J. Pleitez-Navarrete, D. M. Salgado, and J. Brea-Del Castillo, "Congenital and perinatal complications of Chikungunya fever: a Latin American experience," *International Journal of Infectious Diseases*, vol. 51, pp. 85–88, 2016.

[96] "Epidemiological update: Yellow Fever. 20 march 2018," Mar 2018. www.paho.org.

[97] M. Diallo, J. Thonnon, M. Traore-Lamizana, and D. Fontenille, "Vectors of Chikungunya virus in Senegal: current data and transmission cycles.," *The American journal of tropical medicine and hygiene*, vol. 60, no. 2, pp. 281–286, 1999.

[98] "Number of reported cases of Chikungunya fever in the Americas, by country or territory 2017 (to week noted) cumulative cases epidemiological week / ew 51," Dec 2017. www.paho.org.

[99] "WHO - emergencies preparadness, response - Chikungunya," Feb 2018. http://www.who.int/csr/don/archive/disease/chikungunya/en/.

[100] D. J. Gubler, "Dengue/dengue haemorrhagic fever: history and current status," in *New Treatment Strategies for Dengue and Other Flaviviral Diseases: Novartis Foundation Symposium 277*, pp. 3–22, Wiley Online Library, 2006.

[101] D. Gubler and G. Kuno, "Dengue and Dengue hemorrhagic fever: its history and resurgence as a global public health problem, p 1-22," *Dengue and dengue hemorrhagic fever. CAB international, London, United Kingdom*, 1997.

[102] A. Wilder-Smith and D. J. Gubler, "Geographic expansion of Dengue: the impact of international travel," *Medical Clinics*, vol. 92, no. 6, pp. 1377–1390, 2008.

[103] S. Hotta, "Experimental studies on Dengue: I. isolation, identification and modification of the virus," *The Journal of infectious diseases*, pp. 1–9, 1952.

[104] M. G. Guzman, S. B. Halstead, H. Artsob, P. Buchy, J. Farrar, D. J. Gubler, E. Hunsperger, A. Kroeger, H. S. Margolis, E. Martínez, *et al.*, "Dengue: a continuing global threat," *Nature Reviews Microbiology*, vol. 8, no. 12supp, p. S7, 2010.

[105] J. Cardosa, M. H. Ooi, P. H. Tio, D. Perera, E. C. Holmes, K. Bibi, and Z. A. Manap, "Dengue virus serotype 2 from a sylvatic lineage isolated from a patient with Dengue hemorrhagic fever," *PLoS neglected tropical diseases*, vol. 3, no. 4, p. e423, 2009.

[106] E. Wang, H. Ni, R. Xu, A. D. Barrett, S. J. Watowich, D. J. Gubler, and S. C. Weaver, "Evolutionary relationships of endemic/epidemic and sylvatic Dengue viruses," *Journal of virology*, vol. 74, no. 7, pp. 3227–3234, 2000.

[107] N. Vasilakis, J. Cardosa, M. Diallo, A. A. Sall, E. C. Holmes, K. A. Hanley, and S. C. Weaver, "Sylvatic Dengue viruses share the pathogenic potential of urban/endemic Dengue viruses," *Journal of virology*, vol. 84, no. 7, pp. 3726–3728, 2010.

[108] L. Damodaran, A. d. B. Schneider, and D. Janies, "Two ways or one: The relationship of endemic and sylvatic Dengue virus," in *American Journal of Tropical Medicine and Hygiene*, vol. 97, pp. 438–438, ASTMH, 2017.

[109] N. E. A. Murray, M. B. Quam, and A. Wilder-Smith, "Epidemiology of Dengue: past, present and future prospects," *Clinical epidemiology*, vol. 5, p. 299, 2013.

[110] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, *et al.*, "The global distribution and burden of Dengue," *Nature*, vol. 496, no. 7446, p. 504, 2013.

[111] I. Gjenero-Margan, B. Aleraj, D. Krajcar, V. Lesnikar, A. Klobučar, I. Pem-Novosel, S. Kurečić-Filipović, S. Komparak, R. Martić, S. Đuričić, *et al.*, "Autochthonous Dengue fever in Croatia, August-September 2010," *Eurosurveillance*, vol. 16, no. 9, p. 19805, 2011.

[112] J. Lourenço and M. Recker, "The 2012 Madeira Dengue outbreak: epidemiological determinants and future epidemic potential," *PLoS neglected tropical diseases*, vol. 8, no. 8, p. e3083, 2014.

[113] "WHO Western Pacific Region - Dengue situation update number 536," Feb 2018. www.wpro.who.int.

[114] "WHO Dengue fact sheet, 2017," 2017. http://www.who.int/mediacentre/factsheets/fs117/en/

[115] J. E. Staples and T. P. Monath, "Yellow fever: 100 years of discovery," *Jama*, vol. 300, no. 8, pp. 960–962, 2008.

[116] C. S. Bryan, S. W. Moss, and R. J. Kahn, "Yellow fever in the Americas.," *Infectious disease clinics of North America*, vol. 18, no. 2, pp. 275–92, 2004.

[117] P. Nogueira, "The early history of Yellow fever," 2009.

[118] T. P. Monath, "Yellow fever: an update," *The Lancet infectious diseases*, vol. 1, no. 1, pp. 11–20, 2001.

[119] C. L. Gardner and K. D. Ryman, "Yellow fever: a reemerging threat," *Clinics in laboratory medicine*, vol. 30, no. 1, pp. 237–260, 2010.

[120] "WHO Yellow Fever fact sheet," Mar 2016. http://www.who.int/en/news-room/fact-sheets/detail/yellow-fever.

[121] "WHO - Yellow Fever in Brazil," Mar 2018. http://www.who.int/csr/don/09-march-2018-yellow-fever-brazil/en/.

[122] D. Diallo, A. A. Sall, C. T. Diagne, O. Faye, O. Faye, Y. Ba, K. A. Hanley, M. Buenemann, S. C. Weaver, and M. Diallo, "Zika virus emergence in mosquitoes in southeastern Senegal, 2011," *PloS one*, vol. 9, no. 10, p. e109442, 2014.

[123] D. R. Guedes, M. H. Paiva, M. M. Donato, P. P. Barbosa, L. Krokovsky, S. W. dos S Rocha, K. L. Saraiva, M. M. Crespo, T. M. Rezende, G. L. Wallau, *et al.*, "Zika virus replication in the mosquito Culex quinquefasciatus in Brazil," *Emerging microbes & infections*, vol. 6, no. 8, p. e69, 2017.

[124] Z. Liu, T. Zhou, Z. Lai, Z. Zhang, Z. Jia, G. Zhou, T. Williams, J. Xu, J. Gu, X. Zhou, *et al.*, "Competence of Aedes aegypti, Ae. albopictus, and Culex quinquefasciatus mosquitoes as Zika virus vectors, China," *Emerging infectious diseases*, vol. 23, no. 7, p. 1085, 2017.

[125] C. M. Roundy, S. R. Azar, A. C. Brault, G. D. Ebel, A.-B. Failloux, I. Fernandez-Salas, U. Kitron, L. D. Kramer, R. Lourenço-de Oliveira, J. E. Osorio, *et al.*, "Lack of evidence for Zika virus transmission by Culex mosquitoes," *Emerging microbes & infections*, vol. 6, no. 10, p. e90, 2017.

[126] R. Lourenço-de Oliveira, J. T. Marques, V. B. Sreenu, C. A. Nten, E. R. G. R. Aguiar, M. Varjak, A. Kohl, and A.-B. Failloux, "Culex quinquefasciatus mosquitoes do not support replication of Zika virus," *The Journal of general virology*, vol. 99, no. 2, p. 258, 2018.

[127] "Probable Non-Vector-borne Transmission of Zika Virus, Colorado, USA," *Emerging Infectious Diseases*, vol. 17, no. 5, pp. 880–882, 2011.

[128] E. D'Ortenzio, S. Matheron, X. de Lamballerie, B. Hubert, G. Piorkowski, M. Maquart, D. Descamps, F. Damond, Y. Yazdanpanah, and I. Leparc-Goffart, "Evidence of sexual transmission of Zika virus," *New England Journal of Medicine*, vol. 374, no. 22, pp. 2195–2198, 2016.

[129] "Zika suspected and confirmed cases reported by countries and territories in the americas cumulative cases, 2015-2017," Jan 2018. www.paho.org.

[130] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[131] F. Giardina, E. O. Romero-Severson, J. Albert, T. Britton, and T. Leitner, "Inference of transmission network structure from HIV phylogenetic trees," *PLoS computational biology*, vol. 13, no. 1, p. e1005316, 2017.

[132] M. Famulare and H. Hu, "Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria," *International health*, vol. 7, no. 2, pp. 130–138, 2015.

[133] J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, *et al.*, "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples," *nature protocols*, vol. 12, no. 6, p. 1261, 2017.

[134] W. Kermack and A. Mckendrick, "A contribution to the mathematical theory of epidemics," *Proc Roy Soc*, vol. 5, 2003.

[135] V. J. Cook, S. J. Sun, J. Tapia, S. Q. Muth, D. F. Argüello, B. L. Lewis, R. B. Rothenberg, and P. D. McElroy, "Transmission network analysis in tuberculosis contact investigations," *The Journal of infectious diseases*, vol. 196, no. 10, pp. 1517–1527, 2007.

[136] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 295–307, 2005.

[137] B. Facinelli, F. Biavasco, and P. Varaldo, "Use of DNA fingerprinting in an epidemiologic study of outbreak-specific and non-specific strains of group C Neisseria meningitidis," *European journal of epidemiology*, vol. 6, no. 1, pp. 80–83, 1990.

[138] P. D. McElroy, T. R. Sterling, C. R. Driver, B. Kreiswirth, C. L. Woodley, W. A. Cronin, D. X. Hardge, K. L. Shilkret, and R. Ridzon, "Use of DNA fingerprinting to investigate a multiyear, multistate tuberculosis outbreak," *Emerging infectious diseases*, vol. 8, no. 11, p. 1152, 2002.

[139] X. Deng, H. C. den Bakker, and R. S. Hendriksen, "Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens," *Annual review of food science and technology*, vol. 7, pp. 353–374, 2016.

[140] S. McTaggart, C. Nangle, J. Caldwell, S. Alvarez-Madrazo, H. Colhoun, and M. Bennie, "Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies," *International journal of epidemiology*, vol. 47, no. 2, pp. 617–624, 2018.

[141] C. Meaney, R. Moineddin, T. Voruganti, M. A. O'Brien, P. Krueger, and F. Sullivan, "Text mining describes the use of statistical and epidemiological methods in published medical research," *Journal of clinical epidemiology*, vol. 74, pp. 124–132, 2016.

[142] G. Karystianis, I. Buchan, and G. Nenadic, "Mining characteristics of epidemiological studies from Medline: a case study in obesity," *Journal of biomedical semantics*, vol. 5, no. 1, p. 22, 2014.

[143] D. A. Janies, C. Ford, L. Damodaran, and Z. Witter, "Spread of Middle East Respiratory Coronavirus: Genetic versus epidemiological data," *Online journal of public health informatics*, vol. 9, no. 1, 2017.

[144] J. P. Messina, O. J. Brady, T. W. Scott, C. Zou, D. M. Pigott, K. A. Duda, S. Bhatt, L. Katzelnick, R. E. Howes, K. E. Battle, *et al.*, "Global spread of Dengue virus types: mapping the 70 year history," *Trends in microbiology*, vol. 22, no. 3, pp. 138–146, 2014.

[145] N. R. Faria, M. U. G. Kraemer, S. C. Hill, J. Goes de Jesus, R. S. Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C.-H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolay, V. Fonseca, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa,

S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. dos Santos, C. D. dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira, R. G. Abreu, M. H. O. Garcia, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, T. de Oliveira, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, and O. G. Pybus, "Genomic and epidemiological monitoring of Yellow fever virus transmission potential," *Science*, vol. 361, no. 6405, pp. 894–899, 2018.

[146] A. D. Barrett and S. Higgs, "Yellow fever: a disease that has yet to be conquered," *Annu. Rev. Entomol.*, vol. 52, pp. 209–229, 2007.

[147] F. Emmert-Streib, M. Dehmer, and Y. Shi, "Fifty years of graph matching, network alignment and network comparison," *Information Sciences*, vol. 346, pp. 180–197, 2016.

[148] N. Matas, "Comparing network centrality measures as tools for identifying key concepts in complex networks: A case of Wikipedia.," *Journal of Digital Information Management*, vol. 15, no. 4, 2017.

[149] E. C. Holmes, P.-H. Tio, D. Perera, J. Muhi, and J. Cardosa, "Importation and co-circulation of multiple serotypes of Dengue virus in Sarawak, Malaysia," *Virus research*, vol. 143, no. 1, pp. 1–5, 2009.

[150] E. H. Andrade, L. B. Figueiredo, A. P. Vilela, J. C. Rosa, J. G. Oliveira, H. M. Zibaoui, V. E. Araújo, D. P. Miranda, P. C. Ferreira, J. S. Abrahão, *et al.*, "Spatial-temporal co-circulation of Dengue virus 1, 2, 3, and 4 associated with coinfection cases in a hyperendemic area of Brazil: A 4-week survey," *The American journal of tropical medicine and hygiene*, vol. 94, no. 5, pp. 1080–1084, 2016.

[151] B. Mishra, J. Turuk, S. J. Sahu, A. Khajuria, S. Kumar, A. Dey, A. K. Praharaj, *et al.*, "Co-circulation of all four Dengue virus serotypes: First report from Odisha," *Indian journal of medical microbiology*, vol. 35, no. 2, p. 293, 2017.

[152] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?," *Connections (Toronto, Ont.)*, vol. 28, no. 1, p. 16, 2008.

APPENDIX A: ADDITIONAL PHYLOGENETIC TREES

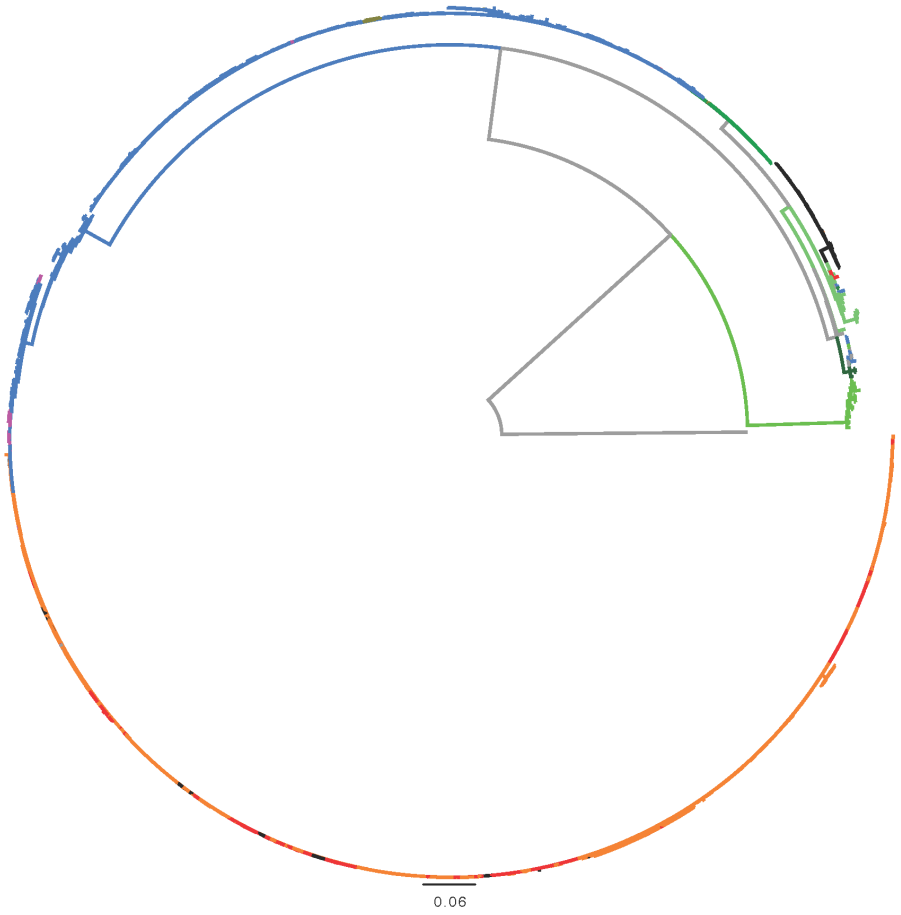

Figure A.1: Maximum-Likelihood phylogenetic tree of 697 Chikungunya virus genomic sequences with branch lengths. Outgroup = O'nyong nyong virus - NC_001512.1.

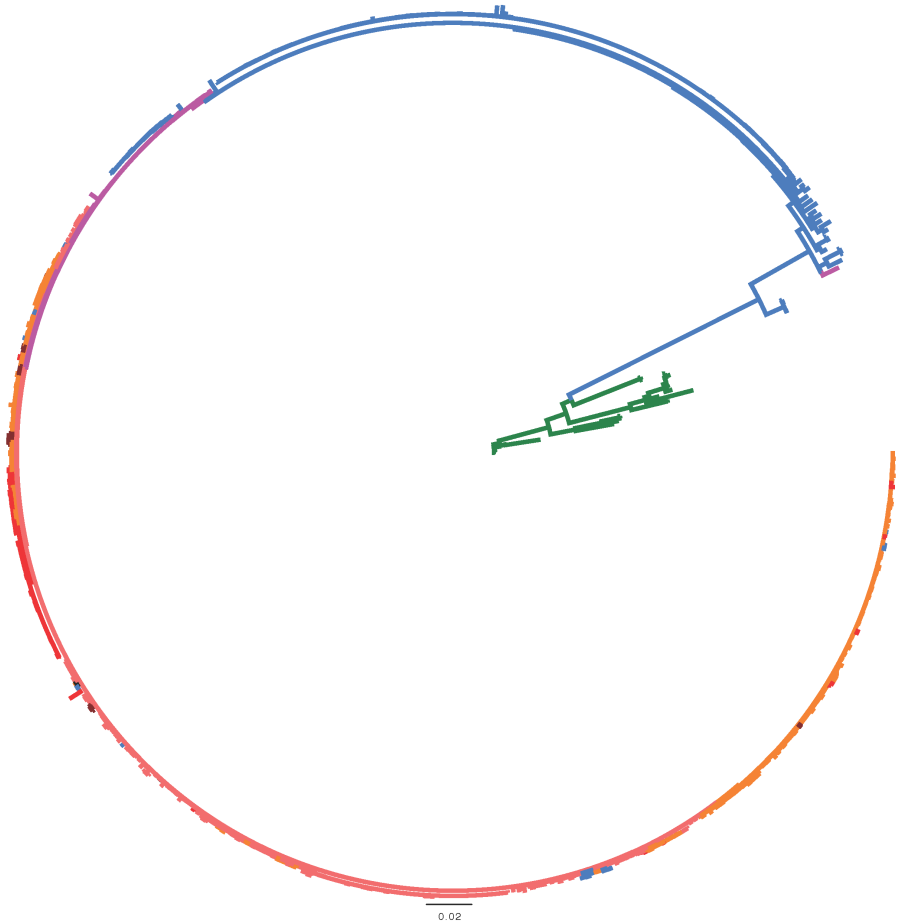

Figure A.2: Maximum-Likelihood phylogenetic tree of 491 Zika virus genomic sequences. Outgroup = LC002520

APPENDIX B: ADDITIONAL MAPPED SYNAPOMORPHIES

Table B.1: Mapped non-ambiguous synapomorphies on nodes of Chikungunya virus phylogenetic trees.

Node	**Lineage**	**N of Mutations**	**Non-ambiguous synapomorphies for each node**
705	South America outbreak sister clade	50	504(C-T); 780(C-T); 1200(T-C); 1934(A-T); 1964(G-A); 1971(C-T); 2059(C-A); 2082(T-C); 2583(C-T); 2649(T-A); 2943(G-A); 3303(C-T); 3342(T-C); 3393(C-T); 4098(G-A); 4149(C-T); 4284(C-T); 4503(G-A); 4878(G-A); 5370(A-T); 5492(C-T); 5700(A-G); 6033(A-T); 6472(C-T); 6507(T-C); 6549(C-T); 6603(A-G); 7059(A-G); 7599(G-A); 7758(T-C); 7879(A-G); 8454(T-C); 8475(T-C); 8889(G-A); 9255(C-T); 9363(G-A); 9672(T-C); 9690(C-T); 10077(C-T); 10123(A-G); 10167(T-C); 11097(C-T); 11355(G-A); 11619(T-C); 11682(A-G); 11929(A-G); 12051(C-T); 12442(T-C); 12474(C-T); 12821(–A)
715	Western Africa	640	216(A-T); 230(A-G); 276(T-C); 441(T-C); 495(G-A); 531(T-C); 618(C-T); 628(A-G); 630(G-A); 651(A-G); 660(T-C); 678(A-G); 681(A-G); 699(A-G); 762(G-A); 780(C-T); 801(C-T); 858(A-G); 867(T-A); 915(C-T); 957(C-T); 966(G-A); 999(A-G); 1032(A-T); 1035(G-A); 1050(G-A); 1104(C-T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	1125(G-T); 1131(A-G); 1143(A-T); 1158(T-C); 1188(G-A); 1197(G-A); 1221(A-G); 1233(C-T); 1239(G-A); 1284(C-T); 1332(G-A); 1368(A-G); 1369(C-A); 1383(C-A); 1392(C-T); 1396(C-T); 1452(C-T); 1477(G-A); 1509(A-T); 1515(A-G); 1518(T-C); 1557(T-C); 1560(C-T); 1575(G-A); 1620(G-A); 1627(T-C); 1632(A-G); 1641(G-A); 1642(A-C); 1663(T-C); 1680(A-G); 1683(C-T); 1687(C-T); 1758(A-G); 1838(T-C); 1953(C-T); 1958(C-A); 1995(G-A); 2007(G-A); 2052(C-T); 2059(C-A); 2062(G-A); 2067(A-T); 2112(A-G); 2145(A-G); 2157(C-T); 2163(G-A); 2179(A-T); 2208(A-G); 2217(T-C); 2235(G-A); 2244(A-G); 2247(C-T); 2262(C-T); 2271(G-A); 2277(A-G); 2283(T-C); 2289(G-A); 2295(A-G); 2301(A-C); 2304(A-G); 2328(T-C); 2364(A-G); 2366(A-G); 2391(A-G); 2412(C-T); 2418(A-G); 2427(T-C); 2433(G-A); 2439(T-A); 2445(C-T); 2472(G-A); 2478(T-C); 2553(T-C); 2577(T-C); 2616(G-A); 2637(A-C); 2685(A-G); 2688(T-A); 2691(T-C); 2697(C-T); 2703(G-A); 2739(C-T); 2748(T-C);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	2853(T-A); 2898(A-G); 2926(T-C); 2931(G-A); 3027(T-C); 3033(A-C); 3051(C-T); 3066(T-C); 3120(G-A); 3135(C-T); 3183(C-T); 3204(G-A); 3243(C-T); 3246(T-C); 3261(T-C); 3267(A-G); 3276(C-T); 3297(T-G); 3306(G-A); 3312(T-C); 3342(T-C); 3369(A-G); 3402(A-G); 3417(A-G); 3435(T-C); 3459(A-G); 3484(T-C); 3498(C-T); 3510(T-C); 3546(G-A); 3564(A-G); 3599(T-C); 3609(G-A); 3624(C-T); 3666(G-T); 3687(C-T); 3705(C-T); 3711(A-G); 3723(A-G); 3726(A-G); 3736(C-T); 3753(C-T); 3762(C-T); 3783(T-C); 3789(T-C); 3792(A-T); 3819(A-G); 3828(A-G); 3870(T-C); 3904(G-A); 3954(A-G); 3984(G-A); 4021(C-T); 4032(A-G); 4056(C-A); 4059(T-C); 4068(C-A); 4069(A-G); 4093(A-C); 4098(G-A); 4150(C-T); 4152(A-G); 4167(C-T); 4170(T-C); 4194(T-C); 4218(A-T); 4227(A-T); 4230(A-C); 4332(C-T); 4455(C-T); 4470(C-T); 4482(T-C); 4491(T-C); 4497(G-A); 4521(T-C); 4575(T-G); 4581(C-T); 4590(T-C); 4599(A-T); 4627(C-T); 4648(A-T); 4650(G-A); 4656(G-A); 4668(C-T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	4674(C-T); 4695(A-G); 4713(T-C); 4752(T-C); 4785(C-A); 4803(C-T); 4812(C-T); 4840(G-T); 4855(C-T); 4857(A-G); 4875(C-T); 4890(T-C); 4899(G-A); 4900(A-G); 4905(T-C); 4923(A-G); 4932(C-T); 4948(C-T); 4950(A-G); 4953(T-C); 4957(C-T); 4989(A-G); 5016(A-G); 5025(G-A); 5064(T-C); 5077(C-A); 5085(C-T); 5124(T-C); 5139(C-T); 5157(A-G); 5166(A-G); 5178(A-G); 5193(C-T); 5226(C-T); 5228(G-A); 5232(T-G); 5246(T-C); 5250(G-A); 5260(A-T); 5324(T-C); 5331(A-T); 5349(C-T); 5352(A-G); 5354(C-T); 5357(T-C); 5367(C-A); 5480(C-T); 5483(A-T); 5719(G-A); 5738(T-A); 5752(G-A); 5753(C-T); 5817(A-G); 5829(G-A); 5849(G-A); 5862(G-C); 5887(A-G); 5907(C-T); 5913(C-T); 5925(G-T); 5929(T-C); 5935(G-A); 5949(A-G); 5957(C-T); 5961(T-C); 5985(A-G); 5988(G-C); 6069(C-T); 6072(C-T); 6084(A-G); 6135(T-C); 6156(T-C); 6174(C-T); 6180(C-T); 6192(G-A); 6204(C-T); 6237(T-C); 6240(T-C); 6243(A-C); 6249(T-C); 6252(A-G); 6260(A-G); 6264(A-T); 6267(A-C);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	6279(G-A); 6291(C-T); 6306(C-T); 6309(C-A); 6354(A-G); 6355(C-T); 6357(A-G); 6358(C-T); 6378(G-A); 6426(A-G); 6445(A-G); 6453(A-G); 6486(A-G); 6501(C-T); 6507(T-C); 6516(C-T); 6528(G-A); 6540(G-A); 6543(T-C); 6579(A-T); 6601(T-C); 6609(A-G); 6618(A-G); 6621(T-A); 6627(A-G); 6645(C-T); 6687(T-C); 6699(A-G); 6714(G-A); 6720(A-G); 6726(G-A); 6732(C-T); 6747(T-C); 6753(C-T); 6762(C-A); 6786(A-G); 6816(A-T); 6834(C-T); 6855(G-A); 6861(A-G); 6864(C-T); 6873(C-T); 6879(A-G); 6888(C-T); 6891(G-T); 6897(T-C); 6909(C-T); 6918(C-T); 6933(A-G); 6948(C-T); 6972(T-C); 6999(G-A); 7002(G-A); 7026(A-C); 7029(A-G); 7044(G-A); 7074(C-T); 7077(A-T); 7095(T-C); 7113(T-G); 7116(A-G); 7158(G-A); 7182(C-T); 7248(A-G); 7275(C-T); 7278(T-C); 7287(A-C); 7314(T-C); 7320(A-G); 7329(C-T); 7344(T-C); 7363(C-T); 7365(T-G); 7458(T-C); 7461(T-C); 7473(T-C); 7476(G-C); 7482(C-T); 7503(A-G); 7620(C-T); 7704(A-G); 7734(T-C); 7750(C-T); 7752(C-T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	7761(G-A); 7762(A-G); 7798(A-C); 7836(C-T); 7845(T-C); 7848(G-C); 7879(A-G); 7914(G-A); 7927(T-C); 7956(A-T); 7992(C-T); 8023(G-A); 8032(T-C); 8040(C-T); 8078(–T); 8101(A-G); 8133(A-G); 8163(G-A); 8166(T-C); 8176(A-G); 8178(T-C); 8190(T-A); 8193(C-T); 8202(C-T); 8269(C-T); 8274(A-C); 8301(A-T); 8307(G-A); 8319(T-C); 8330(A-G); 8334(A-G); 8349(A-G); 8356(A-G); 8359(A-C); 8364(C-G); 8376(G-A); 8383(A-C); 8406(A-G); 8487(A-G); 8489(C-T); 8535(C-T); 8562(C-T); 8572(T-C); 8589(A-G); 8610(C-T); 8640(C-T); 8652(C-T); 8658(T-C); 8679(C-T); 8700(A-G); 8730(T-G); 8733(A-G); 8751(G-A); 8781(G-A); 8811(A-G); 8823(A-T); 8832(A-G); 8877(C-T); 8889(G-A); 8907(T-C); 8920(A-T); 8937(T-C); 8955(C-T); 8999(A-G); 9000(A-C); 9027(C-G); 9064(T-C); 9108(C-T); 9111(C-T); 9138(C-T); 9147(C-T); 9171(C-T); 9189(A-C); 9195(A-G); 9198(A-G); 9219(A-G); 9252(A-G); 9258(A-G); 9327(A-G); 9339(G-A); 9343(T-C); 9351(A-G); 9354(G-T); 9378(A-G);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	9393(C-T); 9408(T-C); 9429(G-A); 9447(T-C); 9459(T-C); 9471(G-A); 9483(C-T); 9486(C-T); 9512(A-G); 9522(C-T); 9528(G-A); 9531(G-A); 9544(C-T); 9579(A-C); 9618(C-T); 9639(A-G); 9675(C-G); 9714(T-C); 9717(A-G); 9723(A-G); 9757(G-A); 9765(A-G); 9777(G-A); 9816(G-A); 9846(A-T); 9870(T-C); 9874(C-T); 9894(G-A); 9966(C-T); 9969(A-G); 9976(C-T); 10017(A-G); 10046(G-C); 10065(A-G); 10080(G-T); 10107(T-C); 10125(A-G); 10140(C-T); 10167(T-C); 10198(G-A); 10201(G-A); 10224(A-G); 10242(T-C); 10244(T-C); 10260(C-T); 10261(A-G); 10272(A-G); 10297(C-T); 10299(G-A); 10311(T-C); 10314(C-T); 10317(C-T); 10323(C-T); 10335(A-G); 10341(C-T); 10354(G-A); 10359(A-G); 10371(C-T); 10386(A-G); 10389(C-T); 10401(G-A); 10411(T-C); 10416(T-C); 10443(A-C); 10452(T-C); 10479(A-G); 10494(A-G); 10498(T-C); 10647(G-A); 10656(T-C); 10680(T-C); 10713(G-C); 10728(A-G); 10788(C-T); 10794(C-A); 10830(C-T); 10884(A-T); 10905(A-G); 10932(A-G); 10935(A-G); 10938(T-G);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	10965(T-C); 10978(T-G); 11154(C-T); 11163(C-T); 11166(G-A); 11175(C-T); 11184(C-T); 11208(G-A); 11214(A-G); 11235(C-T); 11238(G-A); 11346(G-A); 11355(G-A); 11370(A-G); 11373(T-C); 11376(C-T); 11385(A-T); 11388(A-G); 11409(G-A); 11412(C-T); 11436(C-T); 11463(C-T); 11472(C-T); 11478(C-T); 11523(C-T); 11526(T-C); 11592(G-A); 11613(C-T); 11616(T-C); 11628(T-C); 11673(G-A); 11682(A-G); 11694(A-G); 11697(C-T); 11709(C-G); 11739(G-A); 11775(G-A); 11805(C-T); 11808(T-A); 11853(G-A); 11857(C-T); 11860(G-A); 11880(G-A); 11929(A-G); 11956(–A); 11957(–A); 11958(–A); 11960(–T); 11961(–A); 11962(–G); 11963(–A); 11964(–A); 11965(–A); 11966(–G); 11967(–T); 11968(–A); 11969(–C); 11970(–A); 11971(–T); 11972(–A); 11973(–A); 11974(–C); 12172(A-G); 12173(C-A); 12178(G-T); 12179(T-C); 12180(A-T); 12439(–T); 12465(T-C); 12570(C-T); 12854(–G); 12867(–A); 12868(–G); 12869(–T); 12870(–G); 12871(–T); 12872(–G); 12873(–T); 12874(–A); 12875(–C); 12876(–C); 12877(–C);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
715 cont'd	Western Africa	640	12878(–A); 12879(–A); 12880(–A); 12881(–A); 12882(–G); 12883(–A); 12884(–G); 12885(–G); 12886(–T); 12887(–A); 12888(–C); 12889(–A); 12890(–G); 12891(–T); 12892(–A); 12893(–A); 12894(–G); 12895(–A); 12896(–A); 12897(–T); 12920(T-C); 13011(T-C); 13040(A-T); 13060(A–); 13066(G–); 13067(A–); 13068(G–); 13069(A–); 13071(G–); 13072(T–)
719	Indian Ocean Lineage	113	232(G-T); 333(T-A); 414(T-C); 504(C-T); 507(A-G); 660(T-C); 690(G-T); 774(A-T); 831(T-C); 918(A-G); 1029(A-G); 1095(C-T); 1392(C-T); 1704(G-A); 1712(A-G); 1770(T-C); 1853(T-C); 1937(T-C); 1974(C-T); 2018(G-A); 2268(C-T); 2298(T-C); 2466(A-G); 2641(T-C); 3039(G-A); 3141(T-C); 3261(T-C); 3276(C-T); 3357(C-T); 3423(C-T); 3525(G-A); 3660(T-C); 3825(A-G); 3826(C-T); 3954(A-C); 4224(T-C); 4235(C-T); 4458(C-T); 4548(A-G); 4740(G-A); 4875(C-T); 4884(A-T); 4903(T-C); 4935(C-T); 5010(T-C); 5055(C-T); 5079(A-G); 5130(T-C); 5253(A-G); 5264(C-T); 5283(C-T); 5307(C-T); 5367(C-A); 5484(T-C); 5868(C-T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
719 cont'd	Indian Ocean Lineage	113	5988(G-A); 6008(T-C); 6037(C-T); 6127(C-T); 6132(C-A); 6156(T-C); 6439(A-G); 6516(C-T); 6567(C-T); 6723(T-C); 6976(A-G); 7041(G-A); 7050(G-A); 7122(C-T); 7320(A-G); 7461(T-C); 7540(C-T); 7557(T-C); 7665(C-T); 7803(A-T); 7938(T-C); 8098(C-T); 8151(C-T); 8337(G-A); 8364(C-T); 8475(T-C); 8503(T-C); 8970(A-G); 9273(T-C); 9300(C-T); 9366(C-T); 9411(G-A); 9504(C-T); 9633(G-A); 9725(T-C); 10028(C-T); 10216(T-A); 10250(T-C); 10293(C-T); 10384(G-A); 10455(C-T); 10488(T-C); 10611(G-T); 10668(G-A); 10728(A-G); 10851(A-G); 10929(C-T); 11343(A-C); 11454(T-A); 11748(T-C); 11808(T-C); 11865(C-T); 11874(A-C); 11931(C-T); 12014(T-A); 12050(A-G); 12853(C-T); 13011(T-A)
928	South America outbreak	139	154(–G); 155(–G); 156(–C); 157(–T); 159(–C); 160(–G); 161(–T); 162(–G); 166(–G); 167(–A); 168(–C); 169(–A); 170(–C); 435(T-C); 555(C-T); 651(A-G); 849(T-C); 915(C-T); 957(C-T); 1194(T-C); 1638(T-C); 1874(G-T); 2037(T-C); 2172(T-C); 2421(G-A); 2589(A-G); 2688(T-C);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
928 cont'd	South America outbreak	139	2919(C-T); 3030(C-T); 3198(T-C); 3384(C-T); 3445(C-T); 3453(C-T); 3471(G-A); 3537(T-C); 3621(T-C); 3657(C-T); 3672(C-T); 3786(C-T); 3791(C-T); 4092(T-C); 4203(C-T); 4314(C-T); 4428(T-C); 4467(A-G); 4484(C-T); 4488(G-A); 4641(T-C); 4683(C-T); 4744(C-T); 4866(C-T); 4896(G-A); 5016(A-G); 5022(C-T); 5025(C-A); 5064(A-G); 5118(T-C); 5205(G-A); 5242(T-G); 5367(C-T); 5377(G-A); 5717(T-C); 5972(C-T); 6334(C-T); 6344(C-T); 6493(C-T); 6661(C-T); 6753(C-T); 6900(C-T); 7023(C-T); 7035(T-C); 7257(C-T); 7344(T-C); 7377(T-C); 7440(C-T); 7563(A-C); 7773(T-C); 8104(C-T); 8111(C-T); 8126(T-A); 8187(T-C); 8199(C-A); 8232(T-C); 8339(A-G); 8463(C-T); 8473(C-T); 8667(G-A); 8929(T-C); 8937(C-T); 8970(A-T); 8982(C-T); 8987(A-G); 9193(C-T); 9314(T-C); 9321(A-G); 9400(G-A); 9510(A-G); 9546(A-G); 9756(G-A); 9804(T-C); 9858(T-C); 9960(T-C); 9990(T-C); 10011(T-C); 10086(C-T); 10398(C-T); 10488(T-C); 10818(C-A); 10866(T-C); 10989(A-T); 11412(C-T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
928 cont'd	South America outbreak	139	11465(C-T); 11616(T-A); 11682(A-G); 11732(C-T); 11814(C-T); 11821(A-C); 11850(T-C); 11910(T-C); 12442(T-C); 12465(T-C); 12823(A-C); 12847(T-C); 12865(C-T); 13073(–T); 13114(–A); 13115(–T); 13116(–T); 13117(–T); 13118(–T); 13119(–G); 13120(–T); 13121(–T); 13122(–T); 13123(–T); 13125(–T); 13126(–A); 13135(–A); 13136(–T)
955	American Lineage	30	12750(–A); 12751(–C); 12752(–A); 12753(–A); 12754(–A); 12755(–T); 12757(–A); 12758(–G); 12759(–A); 12760(–A); 12761(–G); 12805(–T); 12806(–A); 12807(–G); 12808(–T); 12809(–T); 12810(–C); 12811(–A); 12812(–A); 12813(–A); 12814(–G); 12815(–G); 12816(–G); 12817(–C); 12818(–T); 12819(–A); 12820(–T); 12822(–A); 12823(–A); 12824(–A)
745	Asian Urban / American Lineage	371	232(G-T); 346(C-T); 387(T-C); 528(C-T); 546(A-C); 582(C-T); 609(A-G); 645(C-T); 804(T-C); 807(C-T); 846(A-G); 1007(A-C); 1059(C-T); 1186(C-T); 1464(A-G); 1467(A-G); 1602(G-A); 1609(A-G); 1666(A-C); 1677(A-G); 1707(C-T); 1721(G-A); 1769(G-A); 1880(A-C); 1903(C-T); 1922(A-G);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
745 cont'd	Asian Urban / American Lineage	371	2010(C-T); 2370(C-T); 2415(A-G); 2439(T-C); 2510(C-G); 2562(C-T); 2870(A-T); 2901(A-G); 3063(T-C); 3225(A-G); 3501(C-T); 3588(G-A); 3668(C-T); 3756(C-A); 4107(G-A); 4131(C-T); 4161(C-T); 4257(G-A); 4314(C-T); 4374(A-G); 4410(G-A); 4428(C-T); 4617(G-A); 4686(C-A); 4800(C-T); 4848(C-T); 5061(A-T); 5102(G-A); 5103(C-T); 5142(A-G); 5196(C-T); 5272(C-T); 5289(A-G); 5292(C-T); 5303(T-C); 5304(T-C); 5322(C-T); 5486(G-A); 5832(C-T); 5842(C-A); 5868(C-T); 5936(T-C); 6002(C-T); 6264(A-G); 6343(G-T); 6396(C-T); 6429(C-T); 6484(T-G); 6504(T-G); 6510(C-T); 6518(C-T); 6559(C-T); 6615(T-C); 6744(C-T); 6750(C-T); 6759(C-T); 6762(C-T); 6831(A-G); 6920(A-G); 7104(G-A); 7179(A-G); 7191(G-A); 7216(C-T); 7401(G-A); 7464(A-G); 7494(C-T); 7540(C-T); 7593(G-A); 7722(C-T); 7785(A-G); 7801(C-T); 7822(C-T); 7848(G-A); 7878(G-A); 7905(T-A); 7926(A-G); 7938(T-C); 7961(T-C); 8117(C-T); 8127(C-T); 8172(C-T); 8227(C-A); 8232(T-C); 8301(A-G);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
745 cont'd	Asian Urban / American Lineage	371	8306(G-A); 8351(A-G); 8360(C-T); 8364(C-T); 8379(A-C); 8526(A-G); 8913(C-T); 8929(T-C); 8988(G-A); 8991(A-G); 9018(G-A); 9033(A-C); 9098(C-T); 9192(A-G); 9210(T-C); 9243(C-T); 9445(A-G); 9539(A-G); 9563(T-C); 9621(T-A); 9654(C-T); 9673(G-A); 9675(C-T); 9707(G-A); 9726(A-C); 9864(C-T); 9972(A-G); 10197(A-G); 10243(A-G); 10347(C-T); 10522(A-G); 10644(A-T); 10806(T-C); 10837(G-A); 10911(C-T); 10969(A-G); 11176(A-G); 11218(G-T); 11232(A-G); 11241(A-G); 11454(T-C); 11484(C-A); 11505(C-T); 11547(C-A); 11670(C-T); 11748(T-A); 11802(G-A); 12006(T-C); 12029(–A); 12030(–A); 12031(–G); 12032(–T); 12033(–A); 12034(–T); 12035(–A); 12036(–G); 12037(–A); 12038(–T); 12048(A-G); 12078(C-T); 12160(–A); 12161(–G); 12162(–A); 12163(–A); 12164(–A); 12165(–A); 12166(–C); 12167(–C); 12168(–A); 12169(–G); 12170(–A); 12175(A-G); 12181(C-G); 12182(A-G); 12186(A-G); 12195(C-T); 12392(T-C); 12393(T-C); 12394(G-A); 12398(G-A); 12404(–A); 12405(–A);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
745 cont'd	Asian Urban / American Lineage	371	12406(–G); 12407(–A); 12408(–A); 12409(–T); 12410(–C); 12411(–A); 12412(–A); 12413(–T); 12414(–A); 12456(G-A); 12564(A–); 13002(A-G); 13024(T-C); 13039(C-T); 13074(–C); 13075(–A); 13076(–A); 13077(–A); 13078(–G); 13079(–T); 13080(–G); 13081(–G); 13082(–C); 13083(–T); 13084(–A); 13085(–T); 13088(–A); 13089(–A); 13090(–A); 13091(–A); 13092(–C); 13093(–C); 13094(–C); 13095(–T); 13096(–G); 13097(–A); 13098(–A); 13099(–T); 13100(–A); 13101(–G); 13102(–T); 13103(–A); 13104(–A); 13105(–T); 13106(–A); 13107(–A); 13108(–A); 13109(–A); 13110(–C); 13111(–A); 13112(–T); 13113(–A); 13114(–A); 13115(–A); 13116(–A); 13117(–T); 13118(–T); 13119(–A); 13120(–A); 13121(–T); 13122(–A); 13123(–A); 13124(–G); 13125(–G); 13126(–A); 13127(–T); 13128(–C); 13129(–A); 13130(–A); 13131(–A); 13132(–T); 13133(–G); 13134(–A); 13135(–G); 13136(–T); 13137(–A); 13138(–C); 13139(–C); 13140(–A); 13141(–T); 13142(–A); 13143(–A); 13144(–T); 13145(–T);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
745 cont'd	Asian Urban / American Lineage	371	13146(–G); 13147(–G); 13149(–C); 13150(–A); 13151(–A); 13152(–A); 13153(–C); 13154(–G); 13155(–G); 13156(–A); 13157(–A); 13158(–G); 13159(–A); 13160(–G); 13161(–A); 13162(–T); 13163(–G); 13164(–T); 13165(–A); 13166(–G); 13167(–G); 13168(–T); 13169(–A); 13170(–C); 13171(–T); 13172(–T); 13174(–A); 13175(–G); 13176(–C); 13177(–T); 13178(–T); 13179(–C); 13190(–C); 13191(–T); 13192(–A); 13193(–A); 13194(–A); 13195(–A); 13196(–G); 13197(–C); 13198(–A); 13260(–G); 13261(–C); 13262(–C); 13263(–G); 13264(–A); 13265(–A); 13266(–C); 13267(–T); 13268(–C); 13270(–C); 13271(–T); 13272(–T); 13273(–T); 13274(–G); 13275(–A); 13276(–G); 13277(–A); 13278(–T); 13279(–G); 13280(–T); 13281(–A); 13282(–G); 13283(–G); 13285(–A); 13353(–T); 13355(–G); 13356(–C); 13357(–A); 13358(–T); 13359(–A); 13360(–C); 13361(–C); 13362(–G); 13363(–A); 13364(–A); 13365(–C); 13510(–T); 13511(–C); 13512(–T); 13513(–T); 13514(–C); 13515(–C); 13516(–A);

Table B.1 continued from previous page

Node	Lineage	N of Mutations	Non-ambiguous synapomorphies for each node
745 cont'd	Asian Urban / American Lineage	371	13517(–C); 13518(–A); 13519(–A); 13520(–T); 13521(–T); 13522(–C); 13523(–T); 13524(–C); 13525(–C); 13746(–G); 13747(–T)