

**FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ**

**Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina
Investigativa**

DISSERTAÇÃO DE MESTRADO

**DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA
PARA A GENOTIPAGEM DOS VÍRUS DENGUE, ZIKA,
CHIKUNGUNYA E FEBRE AMARELA**

VAGNER DE SOUZA FONSECA

**Salvador - Bahia
2016**

**FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ**

**Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina
Investigativa**

**DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA
PARA A GENOTIPAGEM DOS VÍRUS DENGUE, ZIKA,
CHIKUNGUNYA E FEBRE AMARELA**

VAGNER DE SOUZA FONSECA

Orientador: Prof. Dr. Luiz Carlos Júnior
Alcântara

Dissertação apresentada ao Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para a obtenção do grau de Mestre.

**Salvador – Bahia
2016**

Ficha Catalográfica elaborada pela Biblioteca do
Instituto Gonçalo Moniz / FIOCRUZ - Salvador - Bahia.

F676d Fonseca, Vagner de Souza
Desenvolvimento de ferramentas de bioinformática para a genotipagem dos vírus
dengue, zika, chikungunya e febre amarela / Vagner de Souza Fonseca. - 2016.
76 f.; 30 cm

Orientador: Dr. Luiz Carlos Júnior Alcântara. Laboratório de Hematologia,
Genética e Biologia Computacional.

Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) –
Fundação Oswaldo Cruz, Instituto Gonçalo Moniz. 2016.

1. Arbovirus. 2. Técnicas de Genotipagem. 3. Filogenia. 4. Mineração de Dados.
I.Título.

CDU 575.833:577.22

"DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA PARA A GENOTIPAGEM DOS VÍRUS DA DENGUE, ZIKA, CHIKUNGUNYA E FEBRE AMARELA."

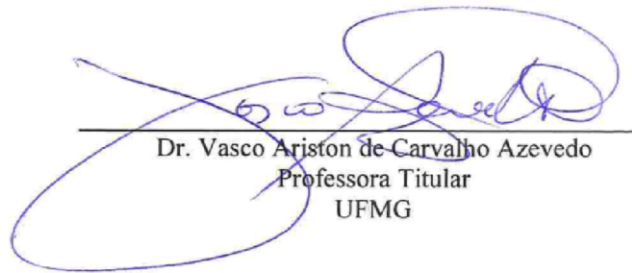
VAGNER DE SOUZA FONSECA

FOLHA DE APROVAÇÃO

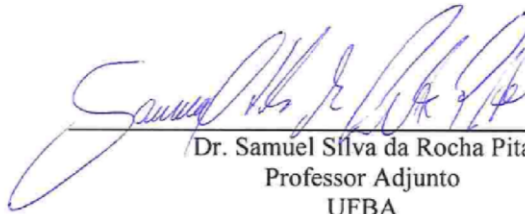
COMISSÃO EXAMINADORA



Dra. Dalila Lucíola Zanette
Pesquisadora
IGM/FIOCRUZ



Dr. Vasco Ariston de Carvalho Azevedo
Professora Titular
UFMG



Dr. Samuel Silva da Rocha Pita
Professor Adjunto
UFBA

AGRADECIMENTOS

Agradeço a Deus, pelo dom da vida e encontrar em suas palavras sabias conforto para superar as minhas dificuldades.

Em segundo lugar, agradeço minha mãe Neide Fonseca, a minha esposa Simara Teixeira por estarem sempre ao meu lado em todos os momentos de minha vida me incentivando e confiando em meu potencial. E todo amor incondicional demonstrado a mim.

Aos meus irmãos Verusca, Tiago e Mariana, aos meus irmãos de coração Aécio e Saulo e aos meus pais de coração Antônio e Teresa, por compartilhar com eles toda a minha trajetória acadêmica e serem compreensivos a todos os momentos.

Agradeço aos grandes professores que ensinaram na pós-graduação, não somente ciência, pois ao longo desses anos tive o prazer de assistir a certas aulas que, com certeza, serão inesquecíveis.

Aos meus colegas do Laboratório de Hematologia, Genética e Biologia Computacional (LHGB), por compartilharem os seus conhecimentos ao longo da jornada do curso, e, mostrando-se sempre dispostos a ajudar um a outro, em especial a Márcia Weber, Marta Giovanetti, Murilo Feire e Inês Restovic.

Ao Dr. Tulio de Oliveira e toda sua equipe do BioAfrica, pelas orientações no desenvolvimento desta pesquisa.

Em especial, agradeço a Luiz Alcântara pela orientação fantástica, ao logo desta pesquisa, e me dar à oportunidade de realizar uma pesquisa interdisciplinar.

Agradeço também aos meus amigos, com quem pude compartilhar conhecimentos e confidências ao longo desses anos.

“A educação é a arma mais poderosa que você pode usar para mudar o mundo”

Nelson Mandela

FONSECA, Vagner de Souza. Desenvolvimento de ferramentas de bioinformática para a genotipagem dos vírus dengue, zika, chikungunya e febre amarela. 76 fl. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) - Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2016.

RESUMO

INTRODUÇÃO: Os Arbovírus transmitidos por mosquitos, como Dengue (DENV), Chikungunya (CHIKV), Zika (ZIKV) e Febre Amarela (YFV), são considerados importantes desafios para a saúde pública. Além do cenário causado pelo DENV, responsável por epidemias há décadas e endêmico em quase todo o país, a introdução do CHIKV e do ZIKV no Brasil traz grande preocupação. Os Arbovírus são transmitidos por mosquitos do gênero *Aedes*, particularmente *Ae. aegypti* e suas doenças relacionadas resultam em aumento dos custos financeiros associados ao diagnóstico e ao tratamento. **MATERIAIS E MÉTODOS:** Para facilitar o diagnóstico e o desenvolvimento de estratégias de prevenção e tratamento de forma eficiente, foram desenvolvidas ferramentas de bioinformática capazes de genotipar esses vírus baseando-se em modelos evolutivos apropriados de forma automática, precisa e rápida. Nesta plataforma, sequências destes arbovírus são selecionadas no *Genbank* por meio de um Sistema Configurável Automático de Mineração (SCAM), para obter um conjunto eficiente de sequências referências que foram utilizadas no desenvolvimento das ferramentas. **RESULTADOS:** Este processo envolveu o alinhamento das sequências referências seguidas por reconstruções de árvores filogenéticas. Para atribuir os genótipos às sequências dos usuários, a ferramenta analisa as sequências uma a uma, através da identificação pelo programa BLAST, seguido pelo alinhamento com o programa ClustalW e posteriormente com a reconstrução filogenética utilizando o programa PAUP*. A classificação genotípica ocorre quando as sequências do usuário se agrupam filogeneticamente com o *bootstrap* igual ou superior a 70%. **CONCLUSÃO:** Essas novas ferramentas de genotipagem automáticas fornecem uma classificação precisa para esses arbovírus mesmo quando as sequências do usuário são oriundas de tecnologias de última geração (NGS), lendo, portanto, fragmentos curtos.

Palavras-Chave: Arbovírus, Técnicas de Genotipagem, Filogenia, Mineração de Dados

FONSECA, Vagner de Souza. Development of bioinformatics tools for the genotyping of dengue, zika, chikungunya and yellow fever viruses. 76 f. il. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) - Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2016.

ABSTRACT

INTRODUÇÃO: Mosquito-borne Arboviruses such as Dengue (DENV), Chikungunya (CHIKV), Zika (ZIKV) and Yellow Fever (YFV) are considered major public health challenges. In addition to the scenario caused by DENV, which has been responsible for epidemics for decades and endemic throughout most of the country, the introduction of CHIKV and ZIKV in Brazil is a major concern. Arboviruses are transmitted by mosquitoes of the genus *Aedes*, particularly *Ae. Aegypti* and its related diseases result in increased financial costs associated with diagnosis and treatment. **MATERIAL AND METHODS:** To facilitate the diagnosis, prevention and treatment strategies efficiently, bioinformatics tools have been developed for the genotyping of these viruses based on appropriate evolutionary models in a automatically, accurately and rapidly manner. In this platform, sequences of these arboviruses are selected in Genbank by means of an Automatic Mining Configurable System (SCAM), to obtain an efficient set of reference sequences that were used in the development of the tools. **RESULT:** This process involved the alignment of the reference sequences followed by phylogenetic tree reconstructions. To assign the genotypes to the user sequences, the tool analyzes the sequences one by one, through identification by the BLAST program, followed by the alignment with the ClustalW program and later with the phylogenetic reconstruction using the PAUP* program. The genotypic classification occurs when the user sequences are grouped phylogenetically with the bootstrap equal to or greater than 70%. **CONCLUSION:** These new automatic genotyping tools provide an accurate classification for these arboviruses even when the user sequences are derived from next-generation technologies (NGS), thus reading short fragments.

Key word: Arboviruses, Phylogeny, Genotyping Techniques, Data Mining

LISTA DE FIGURAS

Figura 1 Reconstrução filogenética do gênero flavivirus. Fonte: adaptado Cook et al., 2012	14
Figura 2 Genoma dos Flavivirus DENV, ZIKV e YFV. Fonte: https://flavivirus.wordpress.com/biosynthesis/	15
Figura 3 Distribuição endêmicas das infecções do YFV no mundo. Fonte: Adaptado CDC, 2016.....	18
Figura 4 Presença das infecções endêmicas causadas pelo DENV no mundo. Fonte: Adaptado CDC, 2016.....	21
Figura 5 Notificações das infecções causadas pelo ZIKV nos últimos 3 meses no mundo. Fonte: CDC 2016	23
Figura 6 Notificações das infecções causadas pelo CHIKV nos últimos 3 meses no mundo. Fonte: CDC 2016	25
Figura 7 Genoma do Vírus Chikungunya. Fonte: PINTO, 2013.....	26
Figura 8 As etapas do processo da Descoberta de Conhecimento de Bases de Dados.	29
Figura 9 Página inicial das ferramentas de bioinformática.	54
Figura 10 Relatório em HTML da ferramenta de identificação dos patógenos contendo informações sobre tipo do vírus, quantidade de sequências, porcentagem da quantidade e uma imagem da legenda do gráfico.	55
Figura 11 Relatório em HTML da ferramenta de genotipagem contendo informações sobre o nome da sequência, comprimento, espécies virais atribuídas, genótipo, e uma ilustração do genoma do vírus.	56
Figura 12 O relatório detalhado em HTML que contém informações como: o nome da sequência, comprimento, vírus atribuído, genótipo, ilustração do genoma viral, análise filogenética, alinhamento e árvore filogenética reconstruída.	57

LISTA DE TABELA

Tabela 1. Análise filogenética de genomas completos realizado ferramenta de genotipagem para classificar o YFV.	58
---	----

LISTA DE ABREVIATURAS E SIGLAS

BD	Banco de dados
BDB	Bancos de Dados Biológicos
BDET	Banco de Dados Especifico Temporário
CHIKV	Vírus Chikungunya (<i>Chikungunya Virus</i>)
CRUD	Criar, Ler, Atualizar e Apagar (<i>Create, Read, Update and Delete</i>)
d.C.	Depois de Cristo
Da	Massa Atômica
datasets	Conjuntos de Dados
DDBJ	Banco de Dados de DNA do Japão (<i>DNA Data Bank of Japan</i>)
DengueDb	Banco de Dados Viral da Dengue (<i>Dengue Viral Database</i>)
DENV	Vírus Dengue (<i>Dengue Virus</i>)
DNA	Ácido Desoxirribonucleico (<i>Deoxyribonucleic Acid</i>)
EBI	<i>Instituto Europeu de Bioinformática (European Bioinformatics Institute)</i>
ECSA	Centro-Leste-Sul Africano (<i>East-Central-South African</i>)
EMBL	<i>Laboratório Europeu de Biologia Molecular (European Molecular Biology Laboratory)</i>
HBV	Vírus Hepatite B (<i>Hepatitis B Virus</i>)
HCV	Vírus Hepatite C (<i>Hepatitis C Virus</i>)
HHV-8	Herpes Vírus Humano Tipo 8 do inglês <i>Human Herpesvirus Type 8</i>
HIV-1	Vírus da Imunodeficiência Humana Tipo 1 (<i>Human Immunodeficiency Virus Type 1</i>)
HIV-2	Vírus da Imunodeficiência Humana Tipo 2 (<i>Human Immunodeficiency Virus Type 2</i>)
HPV	Vírus Papiloma Humano (<i>Human Papilloma Virus</i>)
HTLV-1	Vírus T-Linfotrópico Humano Tipo 1 (<i>Human T-lymphotropic Virus Type 1</i>)
INSDC	<i>International Nucleotide Sequence Database Collaboration</i>
ISF	<i>Insect-Specific Flaviviruses</i>
JEV	Vírus Encefalite Japonesa (<i>Japanese Encephalitis Virus</i>)
KDD	Descoberta de Conhecimento de Bases de Dados (<i>Knowledge Discovery in Database</i>)
MBV	<i>Mosquito-Borne Viruses</i>
NCBI	<i>National Center for Biotechnology Information</i>

NGS	<i>Next-Generation Sequencing</i>
NIH	<i>National Institutes of Health</i>
NKV	<i>No Known Vector Viruses</i>
NS	Proteína Não Estrutural (<i>Nonstructural</i>)
NTF	<i>Non-Translatable Region</i>
ORF	Fase de Leitura Aberta (<i>Open Reading Frame</i>)
RNA	Ácido Ribonucleico (<i>Ribonucleic Acid</i>)
SCAM	Sistema Configurável Automático de Mineração
SGBD	Sistema de Gerenciamento de Bancos de Dados
TBV	<i>Tick-Borne Viruses</i>
UTR	Região não traduzida (<i>Untranslated Region</i>)
WNV	Vírus Oeste do Nilo (<i>West Nile Virus</i>)
YFV	Vírus da Febre Amarela (<i>Yellow Fever Virus</i>)
ZIKV	Vírus Zika (<i>Zika Virus</i>)

SUMÁRIO

1	INTRODUÇÃO	10
1.1	GÊNERO FLAVIVÍRUS	13
1.1.1	Vírus Febre Amarela (YFV)	17
1.1.2	Vírus Dengue (DENV)	19
1.1.3	Vírus Zika (ZIKV)	22
1.2	Vírus Chikungunya (CHIKV)	23
1.3	A MINERAÇÃO DE INFORMAÇÕES PARA A DESENVOLVIMENTO DAS FERRAMENTAS AUTOMATIZADAS	26
1.4	TÉCNICAS DE MINERAÇÃO DE DADOS	28
2	JUSTIFICATIVA	35
3	OBJETIVOS	37
3.1	GERAL	37
3.2	ESPECÍFICOS	37
4	RESULTADOS	38
4.1	ARTIGO	38
4.2	DESENVOLVIMENTO DAS FERRAMENTAS DE IDENTIFICAÇÃO VIRAL, E DA GENOTIPAGEM AUTOMÁTICA DOS DENV, CHIKV, ZIKV E YFV	53
4.3	ANÁLISE DO DESEMPENHO DA FERRAMENTA DO YFV	58
5	DISCUSSÃO	59
6	CONCLUSÕES	61
	REFERÊNCIAS	62
	ANEXO	71

1 INTRODUÇÃO

Os Arbovírus (de “*arthropod borne virus*”) compõem um grupo grande de vírus zoonóticos que infectam artrópodes hematófagos e são comumente transmitidos aos seres humanos, principalmente por meio da picada de mosquitos. Os Arbovírus são classificados em quatro famílias principais: Togaviridae (gênero Alphavirus), Flaviviridae (gênero Flavivirus), Bunyaviridae (gênero Orthobunyavirus e Phlebovírus) e Reoviridae. A maioria dos arbovírus têm um único genoma de ácido ribonucleico (RNA - *Ribonucleic Acid*) de cadeia simples com morfologia esférica e um diâmetro que varia entre 45-120 nanômetro (JAWETZ et al., 2005; FIGUEIREDO et al., 2007).

Estima-se que haja mais de 545 espécies de arbovírus, dentre as quais, mais de 150 estão relacionadas com doenças em seres humanos, sendo a maioria causada por novos agentes ou agentes conhecidos que incidem em locais e espécies que ainda não apresentavam a doença. São mantidos em ciclo de transmissão entre artrópodes (vetores) e reservatórios vertebrados como principais hospedeiros amplificadores (GUBLER, 2001; CLETON et al., 2012).

As manifestações clínicas das arboviroses em seres humanos podem variar desde doença febril, podendo ela ser: indiferenciada, moderada ou grave; erupções cutâneas e artralgia (dor em uma ou mais articulações), a síndrome neurológica e síndrome hemorrágica. A doença febril geralmente se apresenta com sintomas de gripe, como febre, cefaleia, dor retro-orbital e mialgia. A síndrome neurológica pode manifestar-se como mielite, meningite e/ou encefalite, com mudanças de comportamento, paralisia, paresia, convulsões e problemas de coordenação. A artralgia manifesta-se como rash maculopapular ou exantema (erupções cutâneas vermelhas), poliartrite (qualquer tipo de artrite que envolve cinco ou mais articulações) e poliartralgia (dor em varias articulações), enquanto que a síndrome hemorrágica é evidenciada pelas petéquias (pequeno ponto vermelho no corpo), hemorragia e choque combinado com uma redução intensa de plaquetas (CLETON et al., 2012).

1.1 GÊNERO FLAVIVÍRUS

O gênero flavivírus (família do flaviviridae) é composto por 53 espécies diferentes de vírus, abrigando mais de 70 vírus descritos (MUKHOPADHYAY et al., 2005; COOK; HOLMES, 2006; LINDENBACH et al., 2007; COOK et al., 2009). A palavra flavivírus é derivada da palavra *flavus* que tem sua origem no latim e significa amarelo, devido à icterícia (condição que causa uma coloração amarelada da pele) causada pelo Vírus Febre Amarela (YFV - *Yellow Fever Virus*), o protótipo da família (LINDENBACH; RICE, 2001). Em sua grande maioria, são patógenos transmitidos por artrópodes, onde 27 espécies de vírus são transmitidas por mosquito, 12 são transmitidas por carrapato e 14 ainda não possuem seu vetor identificado (GUBLER et al., 2007). Os sintomas da infecção podem alcançar desde febre moderada e mal-estar até encefalite fatal (infecções agudas do encéfalo, causadas por um vírus, bactérias, fungos ou parasitas e até mesmo substâncias químicas ou tóxicas) e febre hemorrágica (GUBKER et al., 2007).

A classificação fundamenta-se em espécies virais considerando a organização genômica, associação dos vetores, morfologia, ecologia viral e a relação das sequências de nucleotídeos. O maior interesse em pesquisas relacionadas aos flavivírus está, principalmente, relacionado ao grande potencial de alguns deles em provocar epidemias associadas a elevadas taxas de mortalidade e morbidade. Os flavivírus que mais infectam hospedeiro humano são os vírus dengue (DENV - *Dengue Virus*), o vírus oeste do Nilo (WNV - *West Nile Virus*), e o vírus encefalite japonesa (JEV - *Japanese Encephalitis Virus*) e o vírus zika (ZIKV - *Zika Virus*) (PIERSON; DIAMOND, 2013).

A construção de análises de inferência filogenéticas com base em alinhamentos de múltiplas sequências de nucleotídeos (nt) dos vírus pertencentes a este gênero indicam a existência de quatro grandes grupos monofiléticos (Figura 1). As sequências codificantes utilizadas na análise deste trabalho de Cook e colaboradores (2002) foram coletadas no *RNA Virus Database* (<http://bioafrica.mrc.ac.za/rnavirusdb/search.php?query=Flavivirus>). O primeiro grupo de sequências, representadas pela cor azul, tem o seu vetor transmitido pelo mosquito, conhecidas na literatura como *Mosquito-Borne Viruses* (MBV). O segundo grupo de sequências, representadas pela cor vermelha, possui o seu vetor transmitido por carrapatos, conhecido na literatura como *Tick-Borne Viruses* (TBV). O terceiro grupo de sequências, representadas pela cor laranja, ainda não possui um artrópode com o vetor conhecido, sendo designadas na literatura como *No Known Vector Viruses* (NKV). O quarto grupo de sequências,

representadas pela cor verde, são sequências transmitidas por insetos específicos, chamadas na literatura de *Insect-Specific Flaviviruses* (ISF). (COOK et al., 2002)

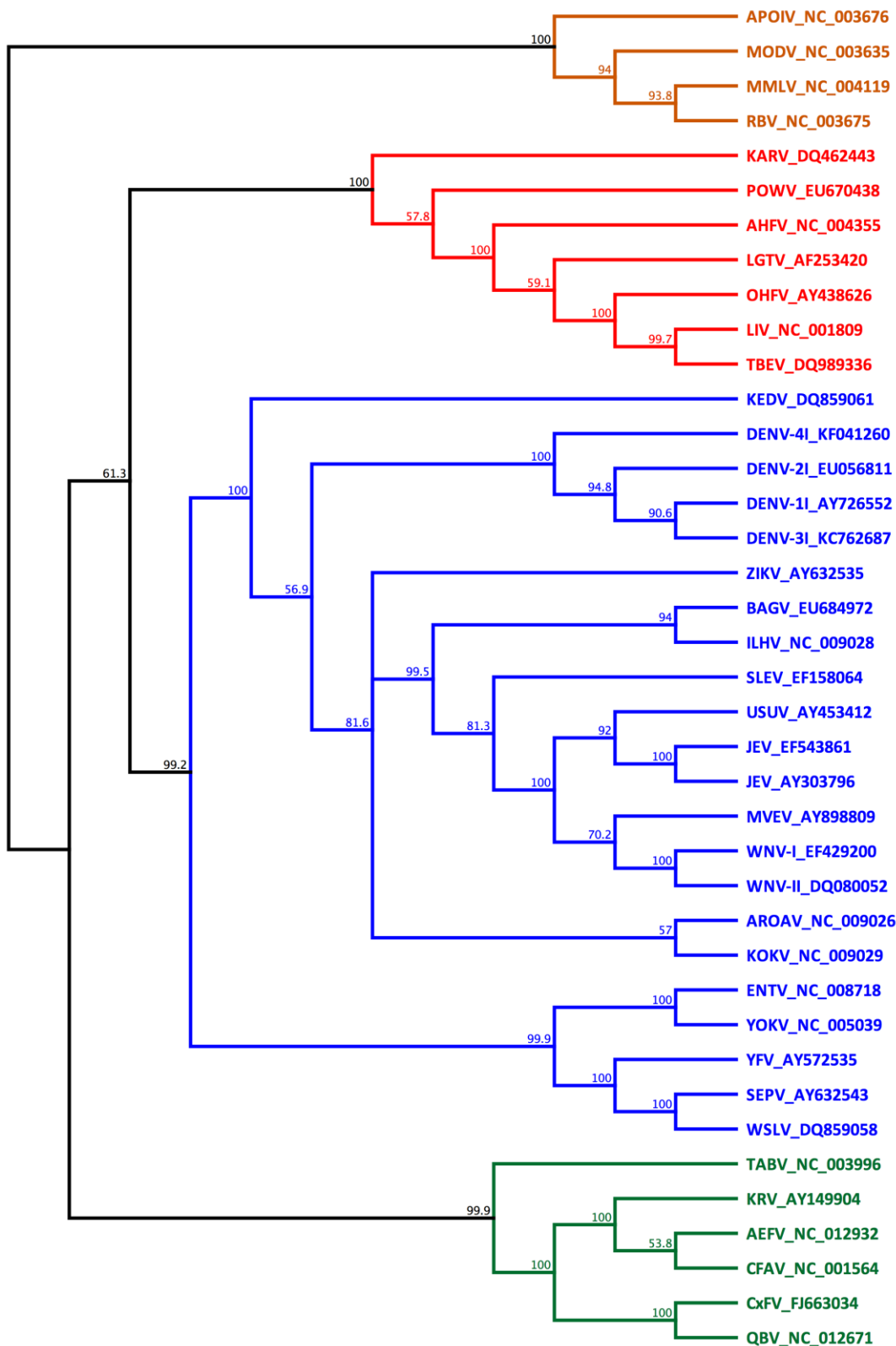


Figura 1. Reconstrução filogenética do gênero flavivirus. Fonte: adaptado Cook et al., 2012

Alguns autores sugerem que o grupo monofilético formado pelo ISF possui uma posição fundamental nas árvores filogenéticas dos flavivírus, constituindo uma linhagem ancestral do gênero (SANG et al., 2003). Os ISF inicialmente foram designados como vírus específicos de mosquitos, porém após a identificação do vírus em flebotomíneos (insetos dípteros e *psychodidae*), esta última nomenclatura foi invalidada e atualmente são chamados apenas de ISF (MOUREAU et al., 2009).

O genoma do gênero flavivírus é composto por uma cadeia simples de RNA com aproximadamente 11kb e com polaridade positiva, que flanqueiam uma única sequência aberta de leitura (ORF - *Open Reading Frame*) codificando uma poliproteína com aproximadamente 3.400 resíduos de aminoácidos. Após sua síntese, esta poliproteína é processada pela proteases virais e celulares com três proteínas estruturais: capsídeo [C], membrana [M] e envelope [E]; e sete proteínas não-estruturais (NS - *Nonstructural*), chamadas de NS1, NS2a, NS2b, NS3, NS4a, NS4b e NS5 (CHAMBERS et al., 1990; SÁNCHEZ-SECO et al., 2005; HARRIS et al., 2006; HOSHINO et al., 2009). Em ambas as extremidades do genoma existem duas regiões não-codificantes (UTR - *Untranslated Region*) denominadas de UTR-5' e UTR-3', as quais tendem a formar estruturas secundárias que desempenham funções reguladoras e de expressão do vírus, como a replicação, virulência e patogenicidade (HOLMES; TWIDDY, 2003) (Figura 2).

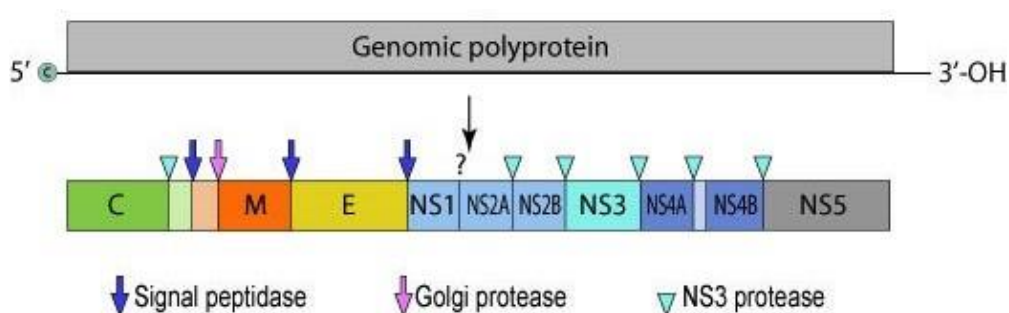


Figura 2. Genoma dos Flavivirus DENV, ZIKV e YFV. Fonte: <https://flavivirus.wordpress.com/biosynthesis/>

O capsídeo possui uma massa atômica molecular com aproximadamente 11kDa, ligando-se fortemente com as moléculas de RNA, e a região C-terminal desta proteína associada ao RNA viral (CHAMBERS et al., 1990a; MA et al., 2004; WANG et al., 2002). A membrana é gerada através da clivagem da proteína prM, que ocorre durante a montagem da partícula viral completa ao longo da passagem pelo complexo de Golgi, originando, assim, partículas virais maduras, possuindo massa atômica de aproximadamente 8kDa (HEINZ; STIASNY, 2012).

O envelope é a maior proteína estrutural, sendo responsável pelas principais propriedades biológicas dos flavivírus, desempenhando um papel importante na patogenicidade em diversas espécies do gênero. A proteína do envelope viral resulta não apenas pela definição do tropismo celular mas também na entrada do vírus na célula alvo (GOLLINS; PORTERFIELD, 1985; REY et al., 1995; MODIS et al., 2004). A proteína do envelope viral compartilha aproximadamente 40% de aminoácidos entre flavivírus dispondo a superfície dos vírion na forma de dímeros (PERERA; KUHN, 2008).

As proteínas não-estruturais são responsáveis pela replicação e transição do RNA viral, no controle da resposta imune no hospedeiro a infecção e no processamento pós-traducional da poliproteína viral (LIU et al., 2004; WESTAWAY et al., 1997). A NS1 é encontrada no citoplasma da célula de mamíferos infectados (WINKLER et al., 1989; FLAMAND et al., 1999). A proteína NS1 partilha um elevado grau de homologia, com 1.056 nucleotídeos que codificam um polipeptídeo de 352 aminoácidos (MACKOW et al., 1987; DEUBEL et al., 1988; MANDL et al., 1989; WRIGHT et al., 1989)

A glicoproteína NS1 possui massa atômica de aproximadamente 46 kDa, contendo 2 ou 3 sítios de glicosilação, e também contém 12 resíduos de cisteína altamente conservados que formam pontes de dissulfeto. Ela é encontrada no interior do retículo endoplasmático, mas também pode ser localizada associada à membrana celular e livre no meio extracelular na forma solúvel. Essa forma solúvel é gerada em quantidades elevadas nas primeiras 48 horas pós-infecção, e é alvo da resposta imune do hospedeiro (CHAMBERS et al., 1990; LINDENBACH; RICE, 2003; LINDENBACH et al., 2007).

As proteínas NS2A, NS2B, NS4A e NS4B são pequenas proteínas hidrofóbicas. A NS2A é uma proteína pequena com massa atômica de aproximadamente 22 kDa, desempenhando um importante papel no processamento da proteína NS1. A NS2B é uma proteína de massa atômica de 14 kDa e associa-se à membrana. A proteína NS4A e a proteína NS4B massa atômica de 16kDa e 27kDa, até o presente momento, não possuem suas funções conhecidas, porem acredita-se que essas proteínas podem ser associadas a replicação agindo como cofatores (LINDENBACH et al., 2007).

A proteína NS3, de aproximadamente 69 kDa, é excessivamente preservada entre os flavivirus e esta associada a funções enzimáticas na replicação e no processamento da poliproteína. As porções N-terminal e C-terminal da proteína NS3 possuem atividades de serina protease na clivagem pós-traducional da poliproteína viral de nucleotídeo de helicase, RNA trifosfatase e trifosfatase (CHAMBERS et al., 1990). Porém as funções do C-terminal não são completamente esclarecidas. Segundo Lindenbach e Rice (2003) a interação entre as proteínas

NS2A e NS3 são especificamente ao domínio de helicase da NS3, para a replicação do RNA viral e a NS2A atuando como cofator.

A proteína NS2B tem semelhança com a ativação da função serina protease da proteína NS3, agindo como um importante cofator (LINDENBACH; RICE, 2003; LINDENBACH et al., 2007). A proteína NS5 possui aproximadamente 103 kDa de massa atômica, sendo a maior proteína não estruturais dos flavivírus, excessivamente preservada, agindo como RNA polimerase RNA dependente localizando-se no citoplasma. A NS5 também apresenta atividade de metiltransferase envolvida na formação do terminal cap 5' do RNA viral (CHAMBERS et al., 1990).

1.1.1 VÍRUS FEBRE AMARELA (YFV)

Estudos demonstram que a origem evolutiva do YFV pertence a África (GOULD et al., 2003; MCNEILL, 2010). Análises filogenéticas indicam que o vírus se originou a partir do Oriente ou África Central, através de transmissão de primatas para humanos, propagando-se para África Ocidental (BRYANT, 2007). O vírus, bem como o vetor *Aedes aegypti*, provavelmente foram trazidos para o Hemisfério Ocidental e para as Américas após a primeira exploração europeia em 1492 com o tráfico de escravos por meio de navios negreiros (HADDOW, 2012).

Os primeiros focos da doença que provavelmente eram da febre amarela ocorreram nas Ilhas de Barlavento do Caribe em Barbados em 1647 e Guadalupe em 1648 (MCNEILL, 2004). De 1640 a 1660, em Barbados, ocorreu o desflorestamento para o cultivo da cana de açúcar pelos holandeses. Este mesmo desflorestamento ocorreu no início do século 18 na Jamaica, Ilha de São Domingos e Cuba para o cultivo da cana de açúcar. Em 1648, colonizadores espanhóis registraram um surto na península de Yucatán, no México, que pode ter sido provocada pelo YFV. Essa doença foi chamada pelo povo Maia de “*Xekik*” (vômito negro) (BRAY, 2004).

Desde o século 17, vários dos principais focos da doença ocorreram nas Américas, África e Europa (WHO, 2014). Nos séculos 18 e 19 a febre amarela era vista como uma das doenças infecciosas mais perigosas (WHO, 2014). Em 1927, o vírus da febre amarela se tornou o primeiro vírus humano a ser isolado (LINDENBACH et al., 2007; SFAKIANOS et al., 2009).

Ocorreram pelo menos 25 grandes surtos na América do Norte. Em 1793, cerca de nove por cento da população da capital dos Estados Unidos foi dizimada pela febre (BRECK, 1929; POWELL, 1949). Em 1878, cerca de 20.000 pessoas morreram em uma epidemia em cidades

nas costas, e dor de cabeça (WHO, 2014). Os sintomas geralmente melhoram dentro de cinco dias (WHO, 2014). Em algumas pessoas, dentro de um dia de melhora, a febre retorna junto com dor abdominal e danos ao fígado que começa a causar a pele amarelada (WHO, 2014). Se isto ocorrer, o risco de problemas de sangramento e do rim é também aumentada (WHO, 2014).

A doença é causada pelo vírus da febre amarela e é transmitida pela picada de um mosquito fêmea infectado (WHO, 2014). Ele infecta apenas humanos, outros primatas, e várias espécies de mosquitos (WHO, 2014). Nas cidades, é transmitido principalmente por mosquitos do tipo *Aedes aegypti* (WHO, 2014). O vírus é um vírus de RNA do gênero Flavivírus (LINDENBACH et al., 2007). A doença pode ser difícil de distinguir de outras doenças, especialmente nas fases iniciais (WHO, 2014). Para confirmar um caso suspeito, o teste de amostra de sangue com a reação em cadeia da polimerase é necessário (TOLLE, 2009).

Os YFV são divididos em quatro genótipos, dois localizados na América do Sul, sendo denominadas de Sul Americano I e Sul Americano II, e dois tipos circulantes no continente africano denominados de Oeste Africano e Leste Africano (VON LINDERN, 2006; DE SOUZA, 2010).

1.1.2 VÍRUS DENGUE (DENV)

O vírus dengue (DENV) causa uma doença infecciosa e debilitante, possuindo 4 sorotipos conhecidos denominados de DENV-1, DENV-2, DENV-3 e DENV-4, tendo como o mosquito *Aedes aegypti* o seu principal transmissor do vetor, a infecção é transmitida para humanos por meio da picada da fêmea do mosquito em regiões tropicais e subtropicais do mundo (GUBLER; CLARK, 1995). Atualmente, a dengue é um problema de saúde pública mundial. São estimados que aproximadamente 3,9 bilhões de pessoas vivem em áreas de risco e que cerca de 390 milhões sejam infectadas por ano por este vetor, resultando 500 mil casos de dengue hemorrágica, levando a óbito 2,5% dos casos de DENV (WHO, 2016). Guble (2002) estimou que devido ao crescimento populacional e o aquecimento global, mais da metade da população mundial viveria em áreas endêmicas ao vetor. O Japão foi o primeiro país a isolar o DENV em seus soldados que estavam com sintomas de febre (KIMURA; HOTTA, 1944). No mesmo período, Sabin e Schlesinger (1945) isolaram outro sorotipo do DENV em soldados americanos.

Os DENV são classificados em quatro sorotipos estreitamente relacionados, denominados: DENV-1, DENV-2, DENV-3 e DENV-4 (HOLMES; TWIDDY, 2003) e 18

genótipos (1-I, 1-II, 1-III, 1-IV, 1-V, 2-I, 2-II, 2-III, 2-IV, 2-V, 2-VI, 3-I, 3-III, 3-V, 4-I, 4-II, 4-III e 4-IV), essa nomenclatura consiste com a anotação do vírus dengue disponível pelo ViPR DB (http://www.viprbrc.org/brc/home.spg?decorator=flavi_dengue).

Assim, os primeiros vetores isolados em soldados japoneses foram denominados de DENV-1 o segundo, isolado em soldados americanos, ficou denominado de DENV-2. Já o DENV-3 e DENV-4 foram isolados nas Filipinas (HAMMON et al., 1960). Documentos chineses publicados durante a dinastia Chin (265 a 420 d.C.) relatam uma doença semelhante a dengue chamada de “veneno da água”. Existem relatos também no período de 1889 a 90, de uma epidemia semelhante ao da dengue em Jakarta, Cairo e Filadélfia (MAIHURU, 2004). Entre as décadas de 50 e 60 do século passado, ocorreu uma epidemia de febre hemorrágica em Manila e Bangkok (HOLMES; TWIDDY, 2003). Antes da Segunda Guerra Mundial, as pandemias ocasionadas pela dengue ocorriam a cada 20 anos, porém não eram frequentes em uma mesma região (SILVA, 2013). Condições propícias decorrentes a mudanças ecológicas e atividades econômicas como a urbanização no sudeste asiático proporcionaram a proliferação do vetor do mosquito, iniciando assim um cenário de pandemia mundial do DENV (GUBLER, 1997; RIGAU-PÉREZ et al., 1998). Devido aos programas de erradicação do vetor *Aedes aegypti* no continente americano, visando o controle da febre amarela nas regiões urbanas das décadas de 50 a 70 do século passado, o vetor da dengue manteve-se localizada nesse período apenas no sudeste asiático com uma grande circulação simultânea de vários sorotipos do vírus da dengue hemorrágica na região (GUBLER, 2002; RIGAU-PÉREZ et al., 1998).

A partir da década de 80, o número de países com epidemia de dengue aumentou significativamente com a introdução de novos genótipos do vírus (RIGAU-PÉREZ et al., 1998). Para Halstead (1997), o mosquito *Aedes aegypti* foi reintroduzido no continente americano com novos sorotipos em populações susceptíveis aumentando assim a transmissão do vetor, pois nas décadas de 60 e 70 havia apenas um único sorotipo circulantes nas américas causando epidemias, em um determinado momento e uma determinada região (WEAVER; VASILAKIS, 2009).

Atualmente a dengue é considerada endêmica em mais de 100 países, distribuídos na Ásia tropical, África, Austrália, América Central e América do Sul, causando altos índices de infecção (Figura 4) (GUBLER, 2002).

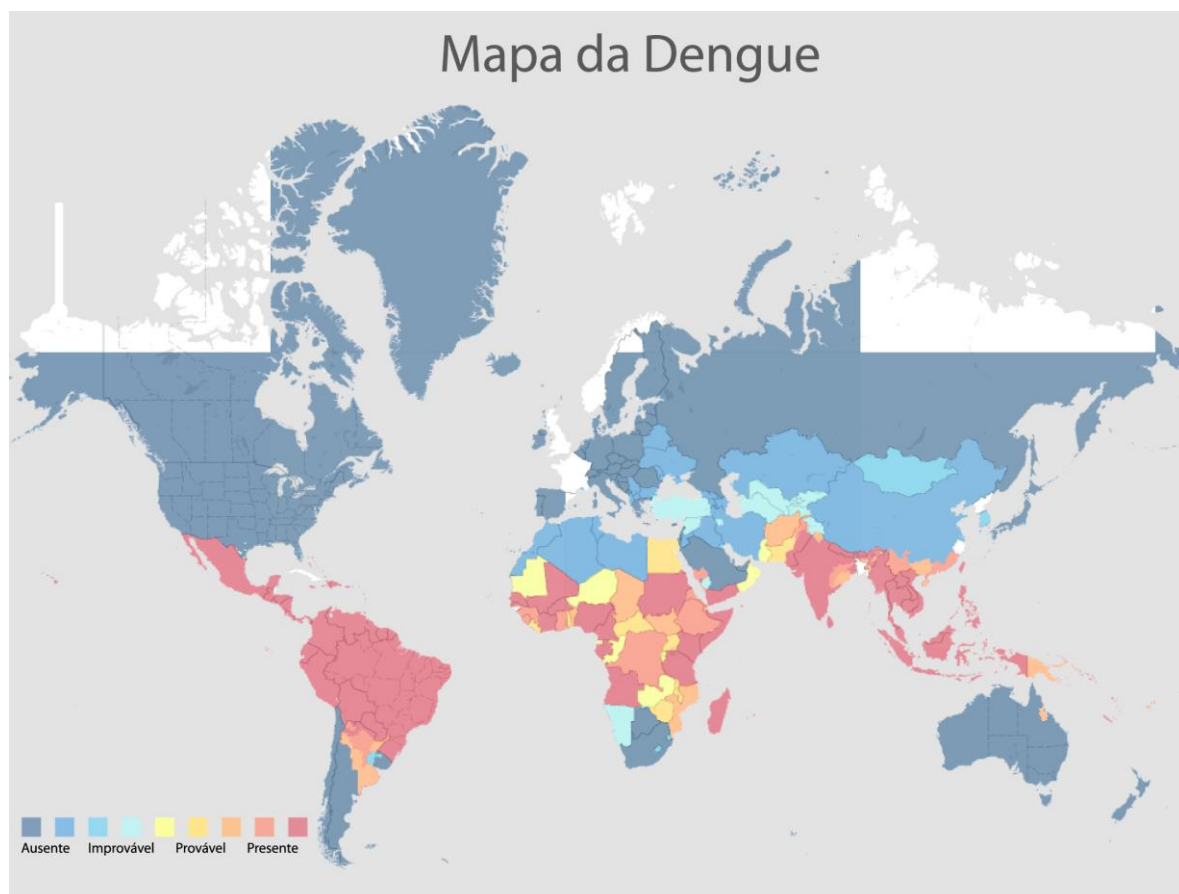


Figura 4. Presença das infecções endêmicas causadas pelo DENV no mundo. Fonte: Adaptado CDC, 2016

A dengue é uma doença tropical e os sintomas começam tipicamente três a quatorze dias após a infecção (KULARATNE, 2015; WHO, 2016a). Isso pode incluir febre alta, dor de cabeça, vômitos, dores musculares e articulares, e uma erupção cutânea característica (WHO, 2016a; KULARATNE, 2015). A recuperação geralmente leva de dois a sete dias (WHO, 2016a). Em uma pequena proporção de casos, a doença evolui para a febre hemorrágica com risco de vida, resultando em sangramento, baixos níveis de plaquetas sanguíneas e extravasamento de plasma sanguíneo, ou em síndrome do choque da dengue, onde a pressão arterial perigosamente baixa ocorre (KULARATNE, 2015).

Atualmente, existem formas de prevenção contra esta patologia. A vacina contra a dengue foi aprovada em três países, mas ainda não está comercialmente disponível (MARON, 2015). A prevenção é reduzindo o habitat do mosquito e limitando a exposição a picadas. Isso pode ser feito por se livrar de ou cobrindo água parada e vestindo roupas que cubram a maior parte do corpo (WHO, 2016a). O tratamento da dengue aguda é de suporte e inclui dar fluido, quer por via oral ou intravenosa para a doença leve ou moderada. Para casos mais graves pode ser necessária transfusão de sangue (KULARATNE, 2015). Cerca de meio milhão de pessoas necessitam de internamento hospitalar por ano (WHO, 2016a).

1.1.3 VÍRUS ZIKA (ZIKV)

O vírus Zika (ZIKV) foi descoberto em Uganda, na Floresta Zika, em um estudo sobre o ciclo do vírus febre amarela. (DICK et al., 1952). A primeira infecção do ZIKV em humanos ocorreu na Nigéria em 1954 (MACNAMARA, 1954). Os sintomas nesta primeira infecção foram febre, dor de cabeça, dor nas articulações difusa, e em um caso, leve icterícia. A primeira infecção detectada pelo vetor *Aedes aegypti* ocorreu na Malásia em 1966 (MARCHETTE et al., 1969). Após 11 anos as primeiras suspeitas de infecções pelo ZIKV, no continente asiático, foi relatado na Indonésia que sete pacientes apresentaram febre, mal-estar, dor de estômago, anorexia e tonturas (OLSON et al., 1981).

Os primeiros surtos da infecção pelo ZIKV ocorreu em uma ilha dos Estados Federados da Micronésia em 2007. Foram identificados em 59 pacientes febre, exantema, conjuntivite e artralgia. Destes, 49 casos foram positivos para ZIKV (DUFFY et al., 2009). Em 2013, ZIKV atingiu a Polinésia Francesa e diversas ilhas da Oceania, onde o *Aedes aegypti* e o *Aedes albopictus* são encontrados em maior parte (HORWOOD et al., 2013). O surto que ocorreu na Polinésia infectou cerca de 10.000 pessoas aproximadamente com sintomas de febre, exantema maculopapular, artralgia e conjuntivite (MUSSO et al., 2014). Em 2014, novos casos também foram registrados na Nova Caledônia e nas Ilhas Cook.

Até a presente data, nenhuma morte foi atribuída ao ZIKV. Em fevereiro de 2014, as autoridades de saúde pública do Chile confirmaram que houve um caso de transmissão autóctone da infecção pelo ZIKV na Ilha de Páscoa (PAHO, 2015). Atualmente, as autoridades de saúde pública do Brasil estão investigando uma possível transmissão do ZIKV, no nordeste do país (PAHO, 2015).

Os recentes surtos de febre zika em diferentes regiões do mundo (Figura 5), demonstram potencial disseminação deste arbovírus em todos os territórios onde os vetores do mosquito *Aedes* estão presentes (CDC, 2016).

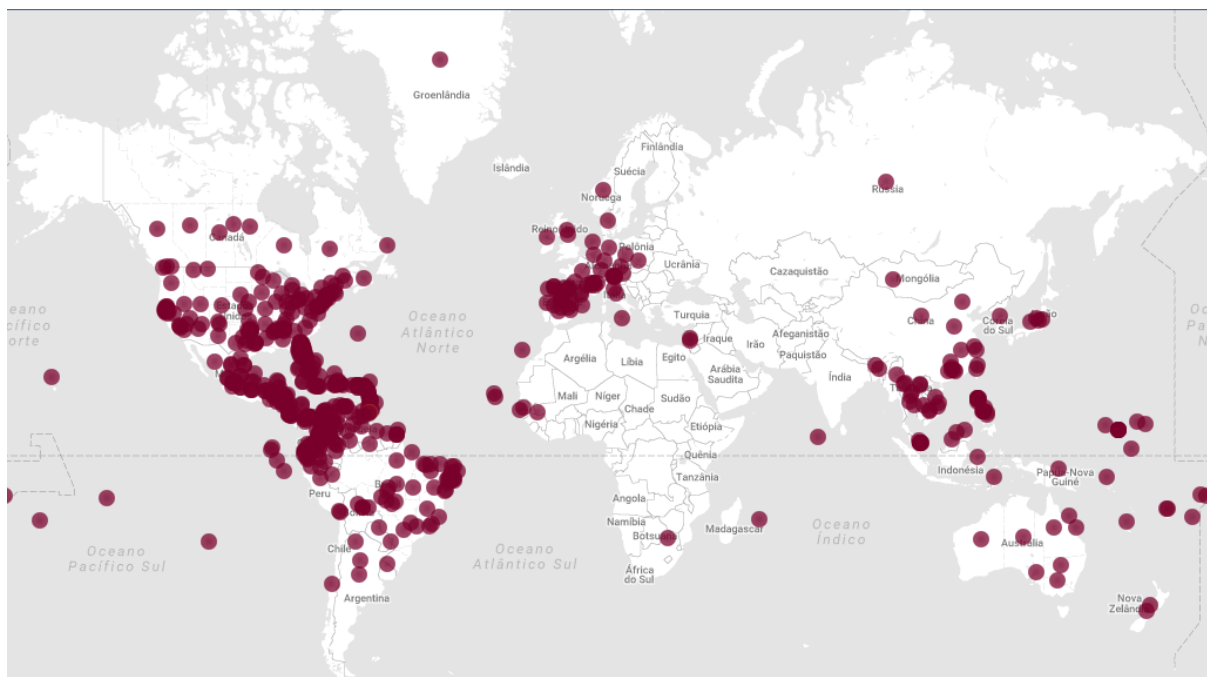


Figura 5. Notificações das infecções causadas pelo ZIKV nos últimos 3 meses no mundo. Fonte: CDC, 2016

Os ZIKV são divididos em três genótipos classificados por análises filogenéticas. Dois deles foram determinados como genótipos africanos, originalmente identificados em um paciente de Uganda em 1947 e em muitos outros países africanos, como o recentemente em Senegal (FAYE et al., 2014). O terceiro grupo foi inicialmente identificado na Malásia em 1966 e tem causado epidemias na Micronésia, Polinésia Francesa e Nova Caledônia (MARCHETTE et al., 1969).

1.2 VÍRUS CHIKUNGUNYA (CHIKV)

O vírus chikungunya (CHIKV), pertence ao gênero *Alphavirus*, da família *Togaviridae*. O CHIKV foi isolado pela primeira vez em amostras de sangue obtidas durante epidemia sugestiva de dengue “*dengue-like*”, ocorrida entre 1952-1953 na Tanzânia, país localizado no sudeste da África. A dificuldade para caminhar provocada pela intensidade do comprometimento das articulações serviu de inspiração para o nome dado àquela doença: chikungunya, que no dialeto Makonde, falado na região significava algo como “andar encurvado sobre o corpo” (ROBINSON, 1995; ROSS, 1956). Nos próximos 50 anos que se

seguiram ao seu isolamento, a circulação do CHIKV estiveram restritas a África e a Ásia (POWERS; LOGUE, 2007).

Em 2004, o CHIKV reemergiu durante epidemia no Quênia, atingindo nos anos seguintes diversas ilhas do Oceano Índico e a Índia. Nesse período, foram identificados centenas de casos importados em países da Europa, Caribe e América do Norte (POWERS; LOGUE, 2007). No ano de 2005, um grande surto de CHIKV ocorreu nas ilhas do Oceano Índico, onde diversos países da Ásia foram afetados em 2006 e 2007, infectando mais de 1,9 milhões de pessoas (WHO, 2016b). Ainda em 2007, a Itália registrou um surto com 197 casos relatados transmitidos pelo vetor do *Aedes albopictus* (WHO, 2016b).

Em outubro de 2013, foram diagnosticados os primeiros casos autóctones na Ilha de Saint Martin, localizada no chamado Caribe Francês (CASSADOU et al., 2014; PAHO, 2015). A partir desse momento a infecção pelo CHIKV foi confirmada em mais de 43 países, sendo esse o primeiro surto documentado de CHIKV nas Américas. Em 2014, cerca de 1.300.000 casos suspeitos de CHIKV foram registrados em ilhas do Caribe, Estados Unidos e países da América Latina, levando a óbito cerca de 191 pessoas (WHO, 2016b).

Atualmente, surtos de CHIKV vem ocorrendo nas Ilhas Cook e Ilhas Marshall, enquanto os casos vêm reduzindo nas Américas. Em 2015, foram notificados aproximadamente 693.489 casos de CHIKV no continente americano, sendo positivo 37.480 casos, destes quase 50% dos casos notificados na América, ocorreu na Colômbia, este numero foi menor que 2014 onde foram registrados mais de 1 milhão de notificações.

A tendência de 2016 é uma queda no número de casos suspeitos, com aproximadamente 31.000 casos notificados até 18 de março de 2016, (Figura 6) representando uma redução de 5 vezes em comparação com o mesmo período de 2015 (WHO, 2016b). Apesar desta tendência, CHIKV continua sendo uma ameaça, principalmente para a Argentina, que registrou o seu primeiro surto de CHIKV.

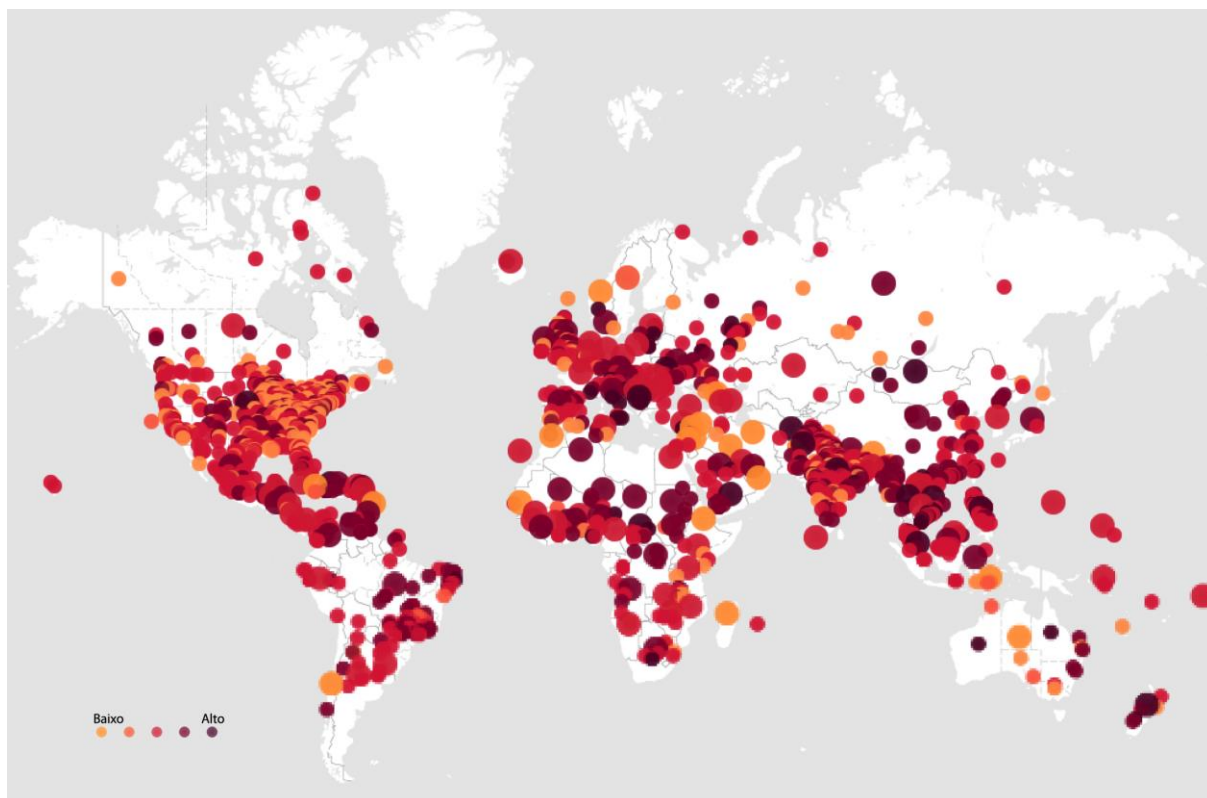


Figura 6. Notificações das infecções causadas pelo CHIKV nos últimos 3 meses no mundo. Fonte: CDC 2016

O genoma do CHIKV consiste em uma molécula de RNA linear, de cadeia simples, polaridade positiva e com aproximadamente 11,8 kb. Este RNA genômico viral assemelha-se aos RNAs mensageiros (mRNAs) celulares por possuir uma estrutura cap na extremidade 5' e uma cauda poli_A na extremidade 3'. O genoma do CHIKV apresenta duas sequências abertas de leitura (ORFs) (Figura 7). Uma delas ocupa os dois terços da porção 5' do genoma e codifica uma poliproteína que, após proteólise, dá origem as proteínas não estruturais (nsP1, nsP2, nsP3 e nsP4) multifuncionais, que formam a replicase viral. A outra ORF, separada da primeira por uma região de junção, codifica uma segunda poliproteína que vai gerar, por processamento proteolítico, as proteínas estruturais [C, E1, PE2 (E3+E2) e 6K]. A região codificante (NTR - *Non-Translatable Region*) é flanqueada nas extremidades 5' e 3' por sequências não traduzidas, denominadas 5'NTR e 3'NTR, respectivamente (LO PRESTI et al., 2012).

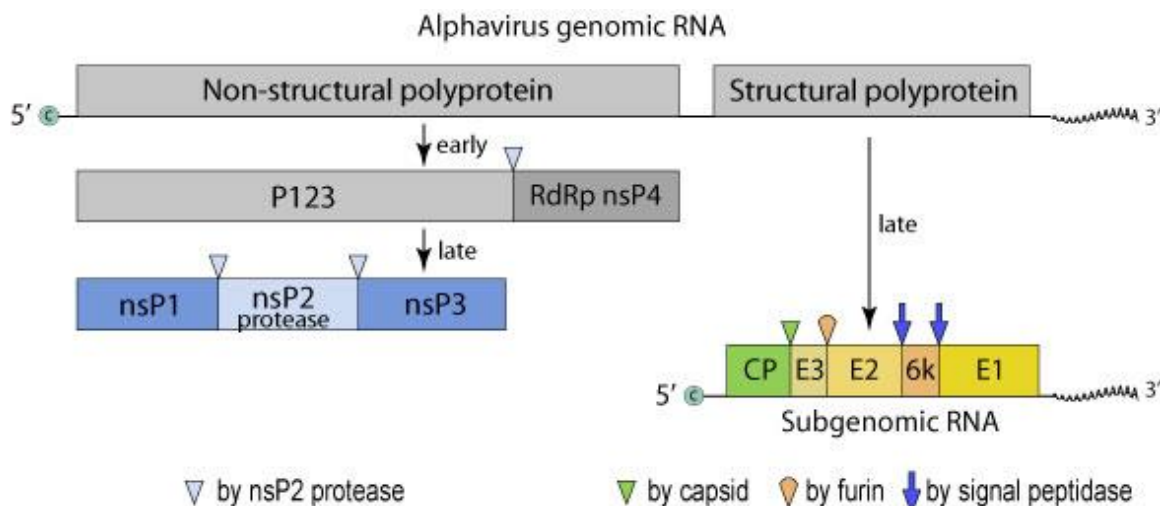


Figura 7. Genoma do Vírus Chikungunya. Fonte: PINTO, 2013

Três grupos distintos do CHIKV, capazes de causar infecção, foram determinados por análise filogenética. O primeiro foi classificado como genótipo Centro-Leste-Sul-Africano (ECSA), originado na África. O segundo genótipo foi classificado como Asiático do Caribe e o terceiro, um genótipo do Oeste Africano, que é mais divergente e menos difundido do que os dois anteriores (VOLK et al., 2010).

1.3 A MINERAÇÃO DE INFORMAÇÕES PARA O DESENVOLVIMENTO DAS FERRAMENTAS AUTOMATIZADAS

A rápida evolução dos recursos computacionais ocorridas nos últimos anos permitiu que fossem geradas grandes volumes de dados armazenados. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses e que o tamanho e a quantidade armazenada nos bancos de dados crescem em uma velocidade maior ainda (DILLY, 1999). O crescimento exponencial desse volume de dados tem gerado uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, a grande massa de dados em informações valiosas. As informações extraídas dos bancos de dados são de grande valia para a tomada de decisões, essas informações na verdade estão implícitas e/ou escondidas sob uma montanha de dados e não podem ser facilmente identificadas utilizando-se sistemas convencionais de gerenciamento de bancos de dados. A partir dessa necessidade surgiu a mineração de dados, denominada de *data mining*.

Um banco de dados (BD) é formado por coleções de informações que se relacionam entre si para criar um sentido. Um Sistema de Gerenciamento de Bancos de Dados (SGBD), é

um conjunto de *softwares* com objetivo de gerenciar o acesso, a manipulação e organização das informações, disponibilizando uma interface para o usuário manipular os dados. Para criação de um banco de dados específico temporário (BDET) será necessária a análise das informações que deverão compô-lo, para a elaboração e execução do seu modelamento.

Os bancos de dados biológicos (BDB) são compostos por tabelas que se relacionam entre si, armazenando uma grande quantidade de registros. As informações contidas neste banco de dados são registros de uma determinada sequência de nucleotídeo, essa sequência normalmente possui uma descrição do nome científico, com as citações na leitura correspondente da sequência. Os BDB possuem a mesma modelagem dos bancos de dados relacionais ou orientados a objetos, apenas o que os caracterizam como biológico são as informações contidas nele. Esses BDB geralmente são associados a um *software* de interface desenvolvido para realização das quatro operações básicas conhecida por CRUD's (*create, read, update, delete*, i.e *insert, select, update e delete*, respectivamente) (BIOINFORMATICS FACTSHEET, 2011).

Uma meta-informação é constituída por características da uma sequência genômica, onde seu objetivo é a tradução dos dados em informações biologicamente importantes. (LEMOS 2004; WEISS, 2010). Essas meta-informações são uma descrição de características em mais alto nível da biossequência. Meta-informações úteis contêm vários tipos de informações, como exemplos, um trecho de ácido desoxirribonucleico (DNA - *deoxyribonucleic acid*) que contém um gene e a sua função (LEMOS, 2004).

Há dois tipos de classificação para BDB, primários e secundários. Os primários são constituídos pela colocação direta de sequências de nucleotídeos, aminoácidos ou estruturas proteicas, sem qualquer processamento ou análise prévia dessas informações. Como exemplo desses BDB podemos citar o banco de dados públicos do *GenBank* pertencente ao *National Center for Biotechnology Information* (NCBI) / *National Institutes of Health* (NIH), *European Bioinformatics Institute* (EBI) *European Molecular Biology Laboratory* (EMBL) e o *DNA Data Bank of Japan* (DDBJ) sob domínio do *International Nucleotide Sequence Database Collaboration* (INSDC). Os secundários são constituídos utilizando informações específicas que são coletadas dos bancos de dados públicos primários (PROSDOCIMI et al., 2002).

Segundo Elmasri e Navathe (2005), os BDB precisam ser principalmente:

- Flexíveis ao lidar com tipos de valores e dados, a colocação de restrições deve ser limitada, uma vez que isso pode excluir valores inesperados, sendo que a exclusão desses valores resulta em perda de informação;

- Fáceis em relação à usabilidade, ou seja, as interfaces do banco de dado devem exibir para os usuários informações de maneira que seja aplicável para o problema que eles estejam tentando tratar e reflita a estrutura dos dados de base;
- Capazes de dar suporte a consultas complexas, pois a definição e a representação destas consultas são extremamente importantes para os estudos biomédicos. Sem conhecimento da estrutura de dados, os usuários comuns não podem construir por conta própria uma consulta complexa através dos dados. Sendo assim, os sistemas devem fornecer ferramentas para que se construam essas consultas.

Devido a que os bancos de dados públicos primários não realizam nenhum tipo de processamento ou análise prévia, as redundâncias e/ou inconsistências das informações são irremissíveis, pois, os laboratórios que alimentam esses bancos possuem critérios particulares sobre a qualidade das sequências a serem publicadas. Com isso, alguns dados armazenados apresentam erros, por possuírem sequências incompletas, corrompidas, e com falhas vindas do próprio sequenciamento, e mesmo assim elas são submetidas a estes bancos de dados.

1.4 TÉCNICAS DE MINERAÇÃO DE DADOS

A mineração de dados (*data mining*) é uma parte do processo conhecido como Descoberta de Conhecimento de Bases de Dados ou *Knowledge Discovery in Database (KDD)*. Este conceito surgiu nos anos 80 para dar vazão ao grande volume de dados que se expandiam exponencialmente, sendo necessário para automatizar a exploração, reconhecimento padrões na modelagem das informações. O KDD é uma tecnologia que surgiu da interseção das áreas da estatística clássica, inteligência artificial e aprendizado de máquina. Segundo Addrians e Zantinge (1996), o KDD permite uma extração não trivial de conhecimento previamente desconhecido e potencialmente útil de um banco de dados. Esse conceito é ressaltado por Fayyad e seus colaboradores (1996) afirmando que a mineração de dados é um processo não trivial de identificações de padrões, desconhecidos, potencialmente úteis e no final das contas, compreensíveis em dados.

Considerando uma hierarquia de complexidade, se algum significado em especial é atribuído a um dado qualquer, esse dado se transforma em uma informação ou fato. Para Sade (1996), se uma norma ou regra é elaborada, a interpretação do confronto entre o fato e a regra constitui em um conhecimento. O processo KDD é constituído de várias etapas, (Figura 8), que são executadas de forma interativa e iterativa. As etapas são interativas porque envolvem a

cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo (BRACHMAN; ANAND, 1996). Por sua vez, a iteração deve-se ao fato de que, com frequência, esse processo não é executado de forma sequencial, mas envolvem repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de *data mining* e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos.

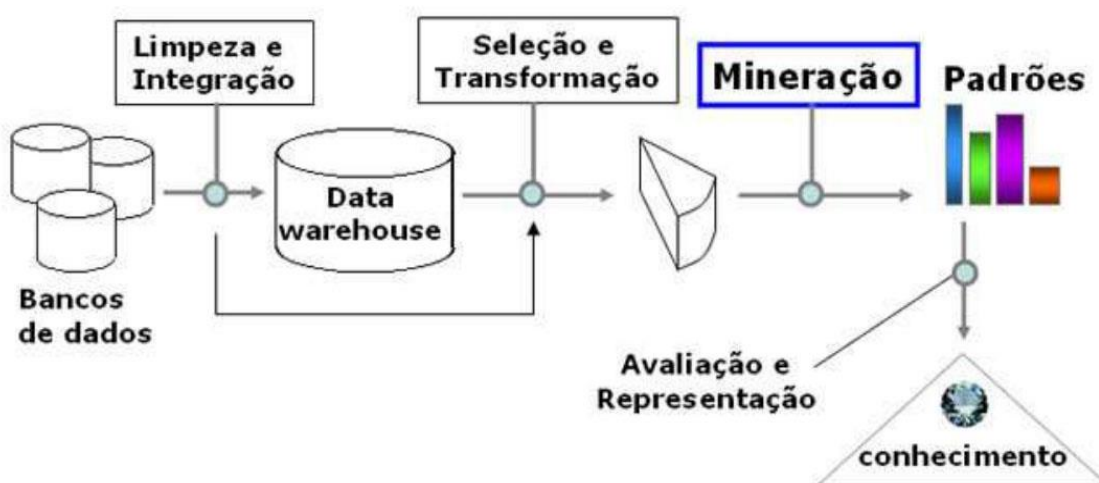


Figura 8. As etapas do processo da Descoberta de Conhecimento de Bases de Dados.

Fonte: <http://www.lsi.ufu.br/documentos/publicacoes/ano/2004/JAI-cap5.pdf>

Esse processo tem início com o entendimento do domínio da aplicação e dos objetivos a serem atingidos. Em seguida, é realizado um agrupamento organizado da massa de dados alvo da descoberta. Como em toda análise quantitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração de dados (DINIZ; LOUZADA-NETO, 2000). A limpeza dos dados, identificada na literatura como *Data Cleaning* é realizada por meio de um pré-processamento, visando assegurar a qualidade dos dados selecionados. Segundo Mannila (1996), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às dificuldades de integração de bases de dados heterogêneas.

Os dados pré-processados devem passar por outra transformação, que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Nessa fase, o uso de *Data Warehouses* expande-se consideravelmente, já que, nessas estruturas, as informações estão alocadas da maneira mais eficiente. O *Data Warehouse* funciona como um depósito central de dados, extraído de dados operacionais, em que a informação é orientada a assuntos, não volátil e de natureza histórica (ADDRIANS; ZANTINGE, 1996). Devido a essas características, o

Data Warehouses tende a se tornar grandes repositórios de dados extremamente organizados, facilitando a aplicação do *Data Mining*.

Prosseguindo no processo KDD, chega-se especificamente à fase de *Data Mining*. O objetivo principal desse passo é a aplicação de técnicas de mineração nos dados pré-processados, o que envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, exige o uso de métodos inteligentes para a extração de padrões ou conhecimentos dos dados.

É importante destacar que cada técnica de *Data Mining* utilizada para conduzir as operações de mineração de dados adapta-se melhor a alguns problemas do que a outros, o que impossibilita a existência de um método de *Data Mining* universalmente melhor. Para cada problema particular, tem-se uma técnica particular.

Portanto, o sucesso de uma tarefa de *Data Mining* está diretamente ligado à experiência e à intuição do analista. A etapa final do processo de mineração consiste no pós-processamento, que engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo identificadas e apresentadas as melhores informações. Dessa forma, o propósito do resultado não consiste somente em visualizar, gráfica ou logicamente, o rendimento da *data mining*, mas, também, em filtrar a informação que será apresentada, eliminando possíveis ruídos, ou seja, padrões redundantes ou irrelevantes que podem surgir no processo.

A mineração de dados pode ser entendida como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa. *Data Mining* define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para prever tendências para o futuro.

Para encontrar respostas ou extrair conhecimento interessante, existem diversos métodos de *Data Mining* disponíveis na literatura. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas. Essas metas são alcançadas por meio dos seguintes métodos de mineração de dados: classificação, modelos de relacionamento entre variáveis, análise de agrupamento, sumarização, modelo de dependência, regras de associação e análise de séries temporais (FAYYAD et al., 1996). É importante ressaltar que a maioria desses métodos é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Essas técnicas vão desde as tradicionais da

estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

Os métodos tradicionais de *Data Mining* são:

- *Classificação*: associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Uma técnica estatística apropriada para classificação é a análise discriminante. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas. A ideia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Para Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, também, a natureza da diferença, de forma que, com base em um conjunto de variáveis independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.
- *Modelos de Relacionamento entre Variáveis*: associa um item a uma ou mais variáveis de predição de valores reais, consideradas variáveis independentes ou exploratórias. Técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y. Observa-se que o método dos mínimos quadrados ordinários, atribuído a Carl Friedrich Gauss, tem propriedades estatísticas relevantes e apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão conforme (GUJARATI, 2000).
- *Análise de Agrupamento (Cluster)*: associa um item a uma ou várias classes categóricas ou *clusters*, em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas.
- Os *clusters* são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos. A análise de *cluster* ou agrupamento é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, o procedimento inicia com o cálculo das distâncias entre

os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles (PEREIRA, 1999). Na sequência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados.

- *Sumarização*: determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são frequentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas, como mínimo, máximo, média, moda, mediana e desvio padrão amostral, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de frequência dos valores. Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas e *box plots*, entre outros. Autores como Levine et al. (2000) e Martins (2001), abordam com grande detalhamento esses procedimentos metodológicos.
- *Modelo de Dependência*: descreve dependências significativas entre variáveis. Modelos de dependência existem em dois níveis: estruturado e quantitativo. O nível estruturado especifica, geralmente em forma de gráfico, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, usando alguma escala numérica. Análises de dependência são aquelas que têm por objetivo o estudo da dependência de uma ou mais variáveis em relação a outras, sendo procedimentos metodológicos para tanto a análise discriminante, a de medidas repetidas, a de correlação canônica, a de regressão multivariada e a de variância multivariada (PADOVANI, 1995).
- *Regras de Associação*: determinam relações entre campos de um banco de dados. A ideia é a derivação de correlações multivariadas que permitam subsidiar as

tomadas de decisão. A busca de associação entre variáveis é, frequentemente, um dos propósitos das pesquisas empíricas. A possível existência de relação entre variáveis orienta análises, conclusões e evidenciação de achados da investigação. Uma regra de associação é definida como *se X então Y*, ou $X \Rightarrow Y$, onde X e Y são conjuntos de itens e $X \cap Y = \emptyset$. Diz-se que X é o antecedente da regra, enquanto Y é o seu consequente. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a frequência de uma regra no universo dos dados minerados. Vários métodos para medir associação são discutidos por Mattar (1998), de natureza paramétrica e não paramétrica, considerando a escala de mensuração das variáveis.

- *Análise de Séries Temporais*: determina características sequenciais, como dados com dependência no tempo. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo. Correlações entre dois instantes de tempo, ou seja, as observações de interesse, são obtidas em instantes sucessivos de tempo, por exemplo, a cada hora, durante 24 horas, ou são registradas por algum equipamento de forma contínua, como um traçado eletrocardiográfico. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares. Há vários modelos estatísticos que podem ser aplicados a essas situações, desde os de regressão linear (simples e múltiplos), os lineares por transformação e regressões assintóticas, além de modelos com defasagem, como os autos regressivos (AR) e outros deles derivados. Uma interessante noção introdutória ao estudo de séries temporais é desenvolvida por Morettin e Tolo (1987).

Diante da descrição sumária de metodologias estatísticas aplicáveis ao procedimento de mineração de dados, registra-se que, embora Hand (1998) afirme que o termo Data Mining possa trazer uma conotação simplista para os estatísticos, (FAYYAD et al., 1996) mostraram a relevância da estatística para o processo de extração de conhecimentos, ao afirmar que essa ciência provê uma linguagem e uma estrutura para quantificar a incerteza resultante quando se tenta deduzir padrões de uma amostra a partir de uma população.

A estatística preocupa-se com a análise primária dos dados, no sentido de que eles são coletados por uma razão particular ou por um conjunto de questões particulares *a priori* (HAND, 1998). *Data Mining*, por outro lado, preocupa-se também com a análise secundária dos dados, num sentido mais amplo e mais indutivo do que uma abordagem hipotético-dedutiva, frequentemente considerada como o paradigma para o progresso da ciência moderna. Assim,

Data Mining pode ser visto como o descendente direto da estatística, já que são técnicas metodológicas complementares.

2 JUSTIFICATIVA

Os DENV, CHIKV, ZIKV e YFV são arbovírus transmitidos principalmente pelo vetor: *Aedes aegypti*. Estes vírus são amplamente distribuídos em regiões tropicais e sub-tropicais (KRAEMER et al., 2015). Nos últimos dois anos, vários estudos têm relatado epidemias simultâneas de DENV, CHIKV, ZIKV e YFV na mesma área geográfica (ROTH et al., 2014; Cardoso et al., 2015).

Apesar do crescente conhecimento sobre estes arbovírus, muitas perguntas permanecem ainda sem respostas sobre seus vetores e reservatórios, sobre a patogênese, diversidade genética, bem como sobre os potenciais efeitos sinérgicos da infecção ou coinfecção entre eles ou com outros vírus circulantes. Estas perguntas destacam a necessidade de pesquisas, principalmente em rede, para otimizar a vigilância, acompanhamento dos doentes e intervenção da saúde pública na atual epidemia causada pelo ZIKV e CHIKV. Dada a rápida propagação do ZIKV e do CHIKV em todo o continente americano, o potencial para complicações neurológicas e uma falta de diagnóstico eficaz, vacina e terapia, estes arbovírus estão sendo vistos como uma questão de saúde pública de grande importância em todas as Américas.

Com o crescente volume de dados gerados por infecções causadas por esses vírus, as pesquisas realizadas sobre o genoma humano, de outras espécies e dos patógenos, acentua-se, ainda mais, a necessidade do uso de sistemas computacionais para auxiliar no processamento deste volume de informação. Neste aspecto pesquisas bioinformáticas processam, em geral, um volume considerável de informações que deve ser analisado em conjunto para proporcionar inferências precisas sobre possíveis hipóteses e possibilitar assim a construção de novas teses sobre o papel ou funcionamento biológico de determinada proteína ou organismo.

Assim, as ferramentas filogenéticas são recursos utilizados no campo da virologia para estudar a evolução viral, traçar a origem de epidemias, estabelecer o modo de transmissão, pesquisar a ocorrência de resistência a medicamentos ou determinar a origem do vírus nos diferentes compartimentos corporais (CHEVENET et al., 2013). Este processo envolve a construção de árvores filogenéticas, sua visualização e interpretação. Sendo assim, a bioinformática é de extrema necessidade para acompanhar a evolução da diversidade viral dando suporte aos estudos de análise de sequências genômicas sendo crucial para a vigilância do polimorfismo viral, no desenvolvimento de novas estratégias terapêuticas, no desenvolvimento de produtos vacinais ou na escolha adequada destes produtos.

Todos os programas de bioinformática, de livre acesso, utilizados para classificação do perfil genético dos subtipos, genótipos, subgrupos ou grupos de vírus **se baseiam no emprego de ferramentas de procura de similaridade** para determinar o genótipo de uma nova sequência, como por exemplo: Dengue Viral Database (DengueDb) do Viral Bioinformatics Resource Center (<http://www.denguedb.org>), NCBI Genotyping Program (<https://www.ncbi.nlm.nih.gov/projects/genotyping/>), entre outros.

Métodos baseados em similaridade são úteis para identificação de padrões de recombinações nas sequências virais, mas eles necessitam de confirmação posterior de métodos filogenéticos próprios e não possuem suporte estatístico para seus resultados. Em contraste, foram desenvolvidas em 2009, sete novas ferramentas de genotipagens virais (HIV-1, HIV-2, HBV, HCV, HTLV-1, HHV-8, e HPV) (<http://www.bioafrica.net/rega-genotype/html>), que são capazes de importar a sequência do usuário, gerar vários segmentos que se sobrepõem bem como vários conjuntos de dados (*datasets*) destes segmentos alinhados com sequências referências (ALCANTARA et al., 2009). **O processamento de vários segmentos com sobreposição ao longo do genoma viral aumenta a precisão e fidelidade dos resultados, especialmente quando se analisa recombinações complexas** (DE OLIVEIRA et al., 2005).

As ferramentas de genotipagem automáticas usam um conjunto de genomas de sequências referência, selecionadas cuidadosamente com a finalidade de representar cada genótipo individual. O uso de uma quantidade de sequências referência que represente o genótipo de um determinado grupo aumenta a consistência e reprodutibilidade dos dados, garantido, assim, uma maior velocidade na busca dos dados e oferecendo um número maior e mais completo de informações evitando que os resultados não sejam limitados por um conjunto inadequado de sequências referências que não represente as informações necessárias para a identificação do vírus.

Assim, propomos o desenvolvimento de uma ferramenta de bioinformática, na plataforma web, capaz de realizar a identificação do genótipo do vírus com a introdução de sequências de nucleotídeos no formato fasta e devolver ao usuário a reconstrução de uma árvore filogenética alinhadas com as sequências referências.

3 OBJETIVOS

3.1 GERAL

Desenvolver ferramentas de bioinformática, na plataforma *web*, capazes de genotipar os vírus Dengue, Zika, Chikungunya e Febre Amarela.

3.2 ESPECÍFICOS

- Desenvolver um *software* na plataforma *web* para permitir a identificação preliminar dos vírus Dengue, Zika, Chikungunya e Febre Amarela;
- Integrar o *software* de identificação preliminar com as ferramentas de bioinformática de genotipagem;
- Disponibilizar na ferramenta árvores filogenéticas, baseadas em análise de *bootstrap*, dos grupos monofilogenéticos das sequências submetidas do usuário, pronta para publicação;
- Desenvolver um algoritmo capaz de testar a eficiência dos resultados da ferramenta de genotipagem.

4 RESULTADOS

Neste capítulo serão apresentados os resultados referentes à criação das 5 ferramentas proposta, principal alvo desta pesquisa. Na sessão 4.1 será apresentado o resultado das ferramentas de Dengue Zika e Chikungunya com uma análise do desempenho da mesma. Os resultados obtidos sobre a o desenvolvimento de ferramentas de genotipagem automatizada para os arbovirus da Dengue, Zika e Chikungunya foram resumidos e organizados em forma de *Sequence Note* intitulado “AN AUTOMATED METHOD FOR THE IDENTIFICATION OF DENGUE, ZIKA AND CHIKUNGUNYA VÍRUS SPECIES AND GENOTYPES”. Na sessão 4.2 será apresentada funcionamento das ferramenta de bioinformática desenvolvidas nesse trabalhos como a de identificação viral, e da genotipagem automática dos DENV, CHIKV, ZIKV e YFV. E por fim na sessão 4.3 estão descritas as análises do desempenho de genotipagem da ferramenta do YFV como descritos no artigo para os demais vírus.

4.1 ARTIGO

An automated method for the identification of Dengue, Zika and Chikungunya virus species and genotypes

Luiz Carlos Júnior Alcântara¹, Nuno R. Faria², Marcio Roberto Teixeira Nunes³, Pieter Libin^{4,5}, Vagner Fonseca¹, Maria Inés Restovic¹, Murilo Freire¹, Marta Giovanetti^{1,6}, Kristof Theys⁵, Lize Cuyper⁵, Ewout Vanden Eynden⁵, Ana Abecasis⁷, Koen Deforche⁸, Gilberto A. Santiago⁹, Isadora Cristina de Siqueira¹, Janaina M. Vasconcelos³, Rivaldo Venâncio da Cunha¹⁰ Oliver G. Pybus², Anne-Mieke Vandamme^{5,7} & Tulio de Oliveira^{1,11*}.

Affiliations

1. Oswaldo Cruz Foundation (FIOCRUZ), Salvador, Bahia, Brazil.
2. Oxford University, UK.
3. Instituto Evandro Chagas, Ananindeua, Para, Brazil.
4. Artificial Intelligence Lab, Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.
5. KU Leuven - University of Leuven, Department of Microbiology and Immunology, Rega Institute for Medical Research, Clinical and Epidemiological Virology, Leuven, Belgium
6. University of Rome “Tor Vergata”.
7. Center for Global Health and Tropical Medicine, Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal.
8. EMWEB solutions for web-based systems, Duigemhofstraat 101, 3020 Herent, Belgium.

9. Dengue Branch, Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, USA.
10. Oswaldo Cruz Foundation (FIOCRUZ), Campo Grande, Mato Grosso do Sul, Brazil.
11. Africa Centre, University of KwaZulu-Natal, Durban, South Africa

* Equal contribution.

Corresponding Author: Prof. Tulio de Oliveira
 Africa Centre for Population Health
 Nelson R Mandela School of Medicine
 University of KwaZulu-Natal, Durban, South Africa
 Email: tdeoliveira@africacentre.ac.za; tuliiodna@gmail.com

Abstract

In recent years, an increasing number of outbreaks of Dengue virus (DENV), Zika virus (ZIKV) and Chikungunya virus (CHIKV) have been reported in Asia and the Americas. The geographical distribution of ZIKV has expanded significantly reported now in at least 41 countries. Since these arboviruses share many clinical symptoms, such as febrile illness with rash, myalgia, or arthralgia, and current serological tests lack the power to discriminate between ZIKV and other flaviviruses, such as DENV and West Nile virus, genetic testing during acute infection has become the standard method to identify the cause of infection. To facilitate diagnosis and the development of prevention and treatment strategies that efficiently target the diversity of these viruses, we developed a rapid high-throughput-genotyping system. The method involves the alignment of a query sequence with a carefully selected set of predefined reference strains, followed by phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequence is assigned the genotype and sub-genotype of the reference strain with the highest bootstrap (>70%) and bootscanning (>90%) scores. The new Arbovirus-Genotyping Tools provide accurate classification of these arboviruses and are currently being assessed for their diagnostic utility.

Introduction

In recent years, an increasing number of outbreaks of *Dengue virus* (DENV), *Zika virus* (ZIKV) and *Chikungunya virus* (CHIKV) have been reported in Asia and the Americas. The geographical distribution of ZIKV has expanded significantly since it was first detected in the Americas in 2015 and ZIKV transmission has now been reported in at least 41 countries.

The same mosquito vector species, *Aedes aegypti* and *Aedes Albopictus* which are widely distributed in tropical and sub-tropical regions (1), mainly transmit DENV, ZIKV and CHIKV. In the past two years, several studies have reported concurrent outbreaks of DENV, ZIKV and CHIKV in the same geographical area (2,3). Currently, unprecedented outbreaks of DENV, ZIKV and CHIKV are co-occurring in Brazil, especially in Bahia and other federal states in the northeast of the country. In 2015, the Brazilian Ministry of Health estimated that approximately 1.6 million cases of DENV, 17,000 cases of CHIKV and between 440,000 and 1.3 million ZIKV cases had occurred in Brazil that year (4). At the end of 2015, a state of national emergency was declared, following suggestions that ZIKV infection might be associated with birth malformations and microcephaly (5). Since then, ZIKV has been detected in at least 41 countries and WHO has formulated a Global Emergency Response Plan. Brazil and many other countries now have concurrent outbreaks of DENV, ZIKV and CHIKV. Since these arboviruses share many clinical symptoms, such as febrile illness with rash, myalgia, or arthralgia, and current serological tests lack the power to discriminate between ZIKV and other flaviviruses, such as DENV and West Nile virus, genetic testing during acute infection has become the standard method to identify the cause of infection. In regions such as Brazil, where ZIKV outbreaks are also concurrent with DENV and CHIKV and there are potential deleterious effects on the fetus in pregnant women, rapid identification of the species and genotype of the replicating virus is crucial. In addition, monitoring virus genotype diversity is critical to understand the emergence and spread of outbreaks, the identification of strains associated with greater epidemic potential, and assessment of genotypes that need to be covered by vaccines. For example, the current worldwide ZIKV outbreak is associated with a genotype of Asian origin (6) and a DENV2 genotype in Nicaragua was shown to have an increased epidemiological fitness that enables more effective evasion of immune responses (7). Further, the CHIKV-ECSA genotype is associated with evolution towards increased transmissibility in regions where the *Aedes albopictus* vector is abundant (8). However, genotypic methods and the genetic regions most suitable for accurate classification remain poorly defined. A genetic variation could be linked to differences in disease severity and/or treatment outcome. Tools for studying the impact of genetic diversity on the biological properties, therapeutic response and epidemic potential of these different arboviruses, remain a major challenge. The new genotyping tools described in this study utilize a sliding window to generate multiple overlapping segments of a query sequence and its reference dataset. Separate phylogenetic trees are reconstructed for each segment, and the reference sequence with the highest bootstrap value is assigned to that segment of the query sequence. Processing of the genome in multiple

segments along the length of the virus increases the accuracy and reliability of the results. Our new genotyping tools use a set of carefully selected full-length reference genomes to represent each individual genotype. The use of multiple reference sequences enhances the consistency and reproducibility of the data and ensures that the phylogenies are not limited by a small number of inappropriate, or uninformative, reference strains.

Methods

To establish the suitability of the genomic regions for classification purposes, we performed Neighborhood Joining, Maximum Likelihood (ML) and Bayesian phylogenetic analyses using previously published whole genome sequences from GenBank: 4,118 of DENV, 63 of ZIKA and 112 of CHIKV.

All selected sequences of DENV, ZIKV and CHIKV were initially aligned and manually edited, then submitted to ML and Bayesian phylogenetic analysis using PhyML (9) and MrBayes (10) to verify the existence of distinct clades corresponding to different serotypes and/or genotypes of DENV, CHIKV and ZIKV.

To evaluate the accuracy and the consistency of identification of viral species, serotype and/or genotype clades using our automated method, an appropriate group of reference sequences were identified (11,12,13). At least ten complete genome sequences per viral genotype and/or serotype were selected. All these reference sequences were re-aligned and submitted to phylogenetic analysis using Neighbor joining (NJ), Maximum Likelihood (ML) and Bayesian methods (14,9,10) to select only those that provided similar topologies using all three different tree-reconstruction methods (Supplementary Table 1).

The selected reference sequences were then evaluated using two different methods with respect to the suitability of sub-genomic regions for automated genotyping. First, a bootscanning sliding window approach was employed by sliding a fixed window size across the full genome. This technique was repeated using increasingly large window sizes ranging from 200 to 2000 nucleotides (nt). All the data sets generated by the sliding windows were used to construct NJ trees considering 1,000 bootstrap replicates to ensure the stability of monophyletic groups.

The second method involved a likelihood-mapping (LM) method using TreePuzzle software (10) to detect the presence of the phylogenetic signal in specific genomic regions of each reference sequence. For DENV and ZIKV, the phylogenetic signal was investigated in the Envelope, NS3 and NS5 regions; for CHIKV, it was investigated in the E1 and E2 regions.

The selected reference sequences were then submitted to a new version of our viral genotyping analysis framework, originally created to classify HIV-1, HCV and HTLV-1 virus sequences

(15). This new version was completely re-coded in JAVA and it is developed as an a free and open source software (FOSS) application in GitHub (<https://github.com/regacev/regagenotype>). The new version of the typing tools accepts up to 20,000 sequences at a time and provide detailed output in XML, CSV and HTML format. This automated method employs BLAST software to identify viral species (DENV, CHIKV, ZIKV) and the genomic region of a query sequence (Figure 1). Identification is accomplished by aligning the query sequence with a codon-aligned complete genome reference dataset, permitting the construction of a phylogenetic tree via PAUP* software using HKY distance methods with gamma distributed rate variation among sites. A given query sequence can thusly be assigned to a particular DENV, CHIKV and ZIKV genotype and/or serotype only if it is shown to cluster monophyletically within a specific group of reference sequences under >70% bootstrap support. To assess the accuracy of our automated method, we compared the genotype/serotype classification results of 4,118 whole genome sequences of DENV, 112 of CHIKV and 63 of ZIKV, to the golden standard ML and Bayesian classification methods (considering a bootstrap value >70% as a cut-off and a posterior probability > 0.9). The sensitivity and specificity of our automated method, both at the first step (species assignment) and at the second step (genotyping), where then calculated. In addition, the automated method's accuracy in classifying sub-genomic regions was also investigated by testing the envelope regions of all of the viral whole genome sequences analyzed.

The products obtained by using non-specific sequencing methods (i.e. short sequence reads produced by Illumina or RNA-seq) were also classified according to genotype/serotype. All published complete genomes of ZIKV, CHIKV and DENV were used to generate short reads 150bp in length across each genome, with a fold coverage of 20 reads per position. A total of 637,602 pseudo-reads for CHIKV, 37,345 for ZIKV and 5,798,596 for DENV were generated. The present study analyzed all published sequences of DENV, CHIKV and ZIKV longer than 150nt, including those collected in Brazil, in order to ensure the applicability of our automated method in the context of outbreak surveillance at a country level.

Results

The phylogenetic analysis conducted by our automated method, employing all of the previously published GenBank whole genome sequences of DENV, ZIKA and CHIKV, correctly classified all the viruses analyzed with respect to genotypes and/or serotype.

For DENV, roman numerals (e.g. I, II, III, etc.) were used to indicate the serotypes identified. A total of 18 subserotypes (1I, 1II, 1III, 1IV, 1V, 2I, 2II, 2III, 2IV, 2V, 2VI, 3I, 3III, 3V, 4I, 4II, 4III and 4IV) were found, which is consistent with the literature. (16)

For CHIKV, three different genotypes were correctly identified: ‘ECSA genotype,’ ‘Asian and Caribbean genotype’ and the ‘West African genotype,’ again in agreement with the literature (17,18).

For ZIKV, three different genotypes were identified in accordance with the literature (11,13,16,12,19): i) The ‘African genotype,’ found in many African countries, discovered in Uganda in 1947; (11) ii) the ‘Asian genotype,’ responsible for the current worldwide epidemic (13,16), originally identified in Malaysia in 1966(12); iii) the ‘divergent West African’ genotype, which was recently identified in Senegal (19).

The results obtained from the bootscanning analysis showed that for ZIKV, the used of segments of around 1,200-1,500bp permit the complete genotyping assignment with a bootstrap value > 70% (Figure 2, Figure 2B). For DENV, only the envelope gene permitted to give a confident genotype assignment across the four different serotypes (Figure 2, Figure 2B). For CHIKV the genotype named ECSA showed to have a poor bootstrap support at the beginning of the genome (non-coding region) around the position 5000 and among the positions 7000 and 9000. Also for CHIV the envelope region ‘E1’ was one that allowed consistent genotyping assignment (Figure 2). (Supplementary Table 2) shows the results obtained analyzing respectively DENV, ZIKV and CHIKV genes.

The likelihood-mapping analysis confirmed that for DENV and ZIKV, the envelope, NS1, NS3 and NS5 genes had a good phylogenetic signal across all serotypes and/or genotypes identified (Supplementary Table 2) and the envelope ‘E2’ gene for CHIKV (Figure 2).

The specificity, sensitivity and accuracy estimated by using our automated method was of 100% for the identification of all the viral species and genotypes (Supplementary Table 3). Only ten among the 4118 DENV whole genome sequences analyzed were not correctly classified at the genotype level, using both the classical phylogenetic analysis and our automated method. Notably, these sequences were outliers in the phylogenetic tree (Supplementary Figure 3), and seven of these sequences (accession numbers KF289073 and AY496879) showed recombination event (Supplementary Figure 4). They were however correctly assigned to a DENV serotype, based on the BLAST results at the first step (Supplementary Table 3).

The classification obtained at the species and at the genotype level was the same using almost all the whole genome sequences selected with a sensitivity >99% and a specificity >99% for all

of the 24 genotypes identified, with the exception of DENV2 genotype 2IV, with a sensitivity of 80.49%, and specificity of 100% (Table 4).

A total of 96.96%, 97.78% and 94.64% pseudo-reads for CHIKV, ZIKV and DENV respectively, were correctly identified, (Figure 5) showing the high curacy with which our method it is also able to classify without any misclassifications the products of non-specific sequencing methods. However, we found that specific genomic regions, like the non-coding regions of ZIKV and CHIKV, and sequences less than 150nt in length, could not be used for species identification.

Overall, our method classified 96% of the whole genome sequences of DENV, CHIKV and ZIKV analyzed at the species level and 86% at the genotypes level. No false positive results were obtained. In detail for DENV, a total number of 186/190 (97.9%) sequences were classified as DENV-1, 208/215 (96.7%) as DENV-2, 373/421 (88.6%) as DENV-3 and 104/104 (100%) as DENV-4. The accuracy at the genotype level was of 171/190 (90%) for DENV-1, 208/215 (96.7%) for DENV-2, 342/421 (81.2%) for DENV-3 and 80/104 (76.9%) for DENV-4 (Supplementary Figure 6).

As expected, using the complete envelope region (1485nt) the accuracy of the Brazilian isolates at the species and genotype level was of 100%. This analysis showed that all 4 DENV serotypes are circulating with varying frequency over time (Supplementary Figure 6).

Our automated method confirmed also the introduction and circulation in Brazil of two CHIKV genotypes, the ECSA and the Asian and Caribbean genotypes (18). ECSA was first introduced in Feira de Santana, Bahia, in June 2014, around the period of the football World Cup, whereas the Asian and Caribbean genotype was introduced in Oiapoque, in the north the country (Table 5). All public ZIKV sequences from the recent outbreak in Brazil, were correctly identified by using our tool as belonging to the Asian genotype.

Identical classification was obtained using a 150nt segment length of the Env gene, the entire envelope gene and the whole genome.

In conclusion our analysis using whole genome sequence of DENV, ZIKV and CHIKV isolated in Brazil and worldwide showed that sub-genomic regions can be correctly classify at the species and genotype level. However, to perform any other kind of analysis at the genotypes level the use of advanced phylodynamic and/or phylogeographic analyses will require the use complete genomes, because these arboviruses seems to have a lower mutation rate than HIV or Influenza virus.

DISCUSSION

The bioinformatics tools introduced provide an accurate and robust framework for the classification of these different arboviruses. By analyzing sequential overlapping segments of a query sequence and its reference alignment, it is possible to construct phylogenetic trees representing each of the segments, conduct bootscanning analyses and draw statistically supported conclusions relating to an arbovirus genotype. The stringent cut-off for bootstrap (>70%) and the bootscan analyses (>90%) greatly reduces the risk of misclassification and the any false-positive results. This enhances the quality of the data.

Conclusion

In conclusion, our new computational method allows the high-throughput classification of DENV, ZIKV and CHIKV species and genotypes in seconds. Species can be classified using short reads from any NGS platform, such as metagenomics Illumina's RNA-seq, and genotypes can be classified most confidently when using envelope gene or complete genome sequences. The framework's is freely available online from a dedicated server (<http://www.bioafrica.net/software.php>).

References:

1. Kraemer, M.U., Sinka, M.E., Duda, K.A., Mylne, A.Q., Shearer, F.M., Barker, C.M., Moore, C.G., Carvalho, R.G., Coelho, G.E., Van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I.R., Teng, H.J., Brady, O.J., Messina, J.P., Pigott, D.M., Scott, T.W., Smith, D.L., Wint, G.R., Golding, N. and Hay, S.I. (15) The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *Elife.*, 4, 08347.
2. Cardoso, C.W., Paploski, I.A., Kikuti, M., Rodrigues, M.S., Silva, M.M., Campos, G.S., Sardi, S.I., Kitron, U., Reis, M.G. and Ribeiro, G.S. (2015) Outbreak of Exanthematous Illness Associated with Zika, Chikungunya, and Dengue Viruses, Salvador, Brazil. *Emerg Infect Dis.*, 21, 2274-6.
3. Roth, A., Mercier, A., Lepers, C., Hoy, D., Duituturaga, S., Benyon, E., Guillaumot, L. and Souares, Y. (2014) Concurrent outbreaks of Dengue, Chikungunya and Zika virus infections - an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012-2014. *Euro Surveill.*, 16, 19-41.
4. Boletim Epidemiológico Secretaria de Vigilância em Saúde (SVS) – Ministério da Saúde, Brazil. 2015, Vol 46, N 42 ISSN 2358-9450 accessible at

<http://u.saude.gov.br/images/pdf/2015/dezembro/11/svs-be-2015-047-dengue-se47-final.pdf>.

5. PAHO/WHO. (2015) Epidemiological Alert. Neurological syndrome, congenital malformations, and Zika virus infection. Implications for public health in the Americas.
6. Faria, N.R., Azevedo, R.D., Kraemer, M.U., Souza, R., Cunha, M.S., Hill, S.C., Thézé, J., Bonsall, M.B., Bowden, T.A., Rissanen, I., Rocco, I.M., Nogueira, J.S., Maeda, A.Y., Vasami, F.G., Macedo, F.L., Suzuki, A., Rodrigues, S.G., Cruz, A.C., Nunes, B.T., Medeiros, D.B., Rodrigues, D.S., Nunes, Queiroz, A.L., Silva, E.V., Henriques, D.F., Travassos, da Rosa, E.S., de Oliveira, C.S., Martins, L.C., Vasconcelos, H.B., Casseb, L.M., Simith, D.B., Messina, J.P., Abade, L., Lourenço, J., Alcantara, L.C., Lima, M.M., Giovanetti, M., Hay, S.I., de Oliveira, R.S., Lemos, P.D., Oliveira, L.F., de Lima, C.P., da Silva, S.P., Vasconcelos, J.M., Franco, L., Cardoso, J.F., Vianez-Júnior, J.L., Mir, D., Bello, G., Delatorre, E., Khan, K., Creatore, M., Coelho, G.E., de Oliveira, W.K., Tesh, R., Pybus, O.G., Nunes, M.R. and Vasconcelos, P.F. (2016) Zika virus in the Americas: Early epidemiological and genetic findings. *Science.*, 24.
7. Manokaran, G., Finol, E., Wang, C., Gunaratne, J., Bahl, J., Ong, E.Z., Tan, H.C., Sessions, O.M., Ward, A.M., Gubler, D.J., Harris, E., Garcia-Blanco, M.A. and Ooi, E.E. (2015) Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. *Science.*, 350, 217-21.
8. Tsetsarkin, K.A. and Weaver, S.C. (2011) Sequential adaptive mutations enhance efficient vector switching by Chikungunya virus and its epidemic emergence. *PLoS Pathog.*, 7, 100-2412.
9. Guindon S, Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52, 696–704.
10. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572–1574.
11. Dick, G.W., Kitchen, S.F. and Haddow, A.J. (1952) Zika virus. I. Isolations and serological specificity. *Trans R Soc Trop Med Hyg.*, 46, 509-20.
12. Marchette, N.J., Garcia, R. and Rudnick, A. (1969) Isolation of Zika virus from *Aedes aegypti* mosquitoes in Malaysia. *Am J Trop Med Hyg* 18, 411–415.
13. Cao-Lormeau, V.M., Roche, C., Teissier, A., et al. Zika virus, French Polynesia, South Pacific, 2013. *Emerg Infect Dis* 2014;20:1085-6.
14. Wilgenbusch J.C., Warren D.L., Swofford D.L. 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.

15. Alcantara, L.C., Cassol, S., Libin, P., Deforche, K., Pybus, O.G., Van Ranst, M., Galvao-Castro, B., Vandamme, A.M. and de Oliveira, T. (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res.*, 37, 634-42.
16. Holmes, E.C. and Twiddy, S.S. (2003) The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol.*, 3, 19-28.
17. Volk, S.M., Chen, R., Tsetsarkin, K.A., Adams, A.P. and Garcia, T.I. (2010) Genome scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. *J Virol.*, 84, 6497-6504.
18. Nunes, M.R., Faria, N.R., de Vasconcelos, J.M., Golding, N., Kraemer, M.U., de Oliveira, L.F., Azevedo, R.S., da Silva, D.E., da Silva, E.V., da Silva, S.P., Carvalho, V.L., Coelho, G.E., Cruz, A.C., Rodrigues, S.G., Vianez, J.L. Jr., Nunes, B.T., Cardoso, J.F., Tesh, R.B., Hay, S.I., Pybus, O.G. and Vasconcelos, P.F. (2015) Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med.*, 13, 102.
19. Faye O, Freire CCM, Lamarino A, Faye O, de Oliveira JVC, et al. (2014) Molecular Evolution of Zika Virus during its Emergence in the 20th Century. *PLoS Negl Trop Dis* 8(1): e2636. doi:10.1371/journal.pntd.0002636.

Figures legends

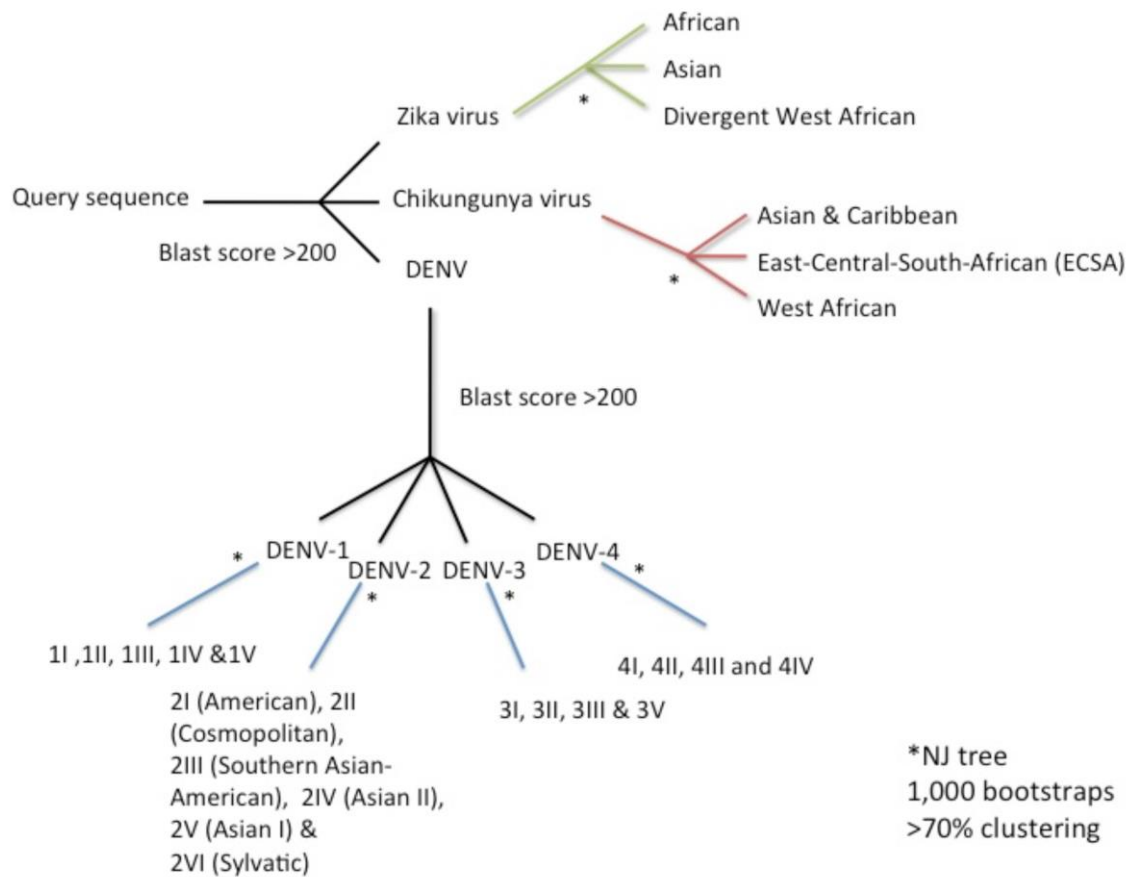


Figure 1: ZIKV, CHIKV and DENV genotyping process. In total, three genotypes are identifiable for ZIKV and CHIKV. For DENV, four serotypes (DENV1-4) and 18 genotypes can be identified. The steps in the assignment to these genotypes by the automated method are also indicated.

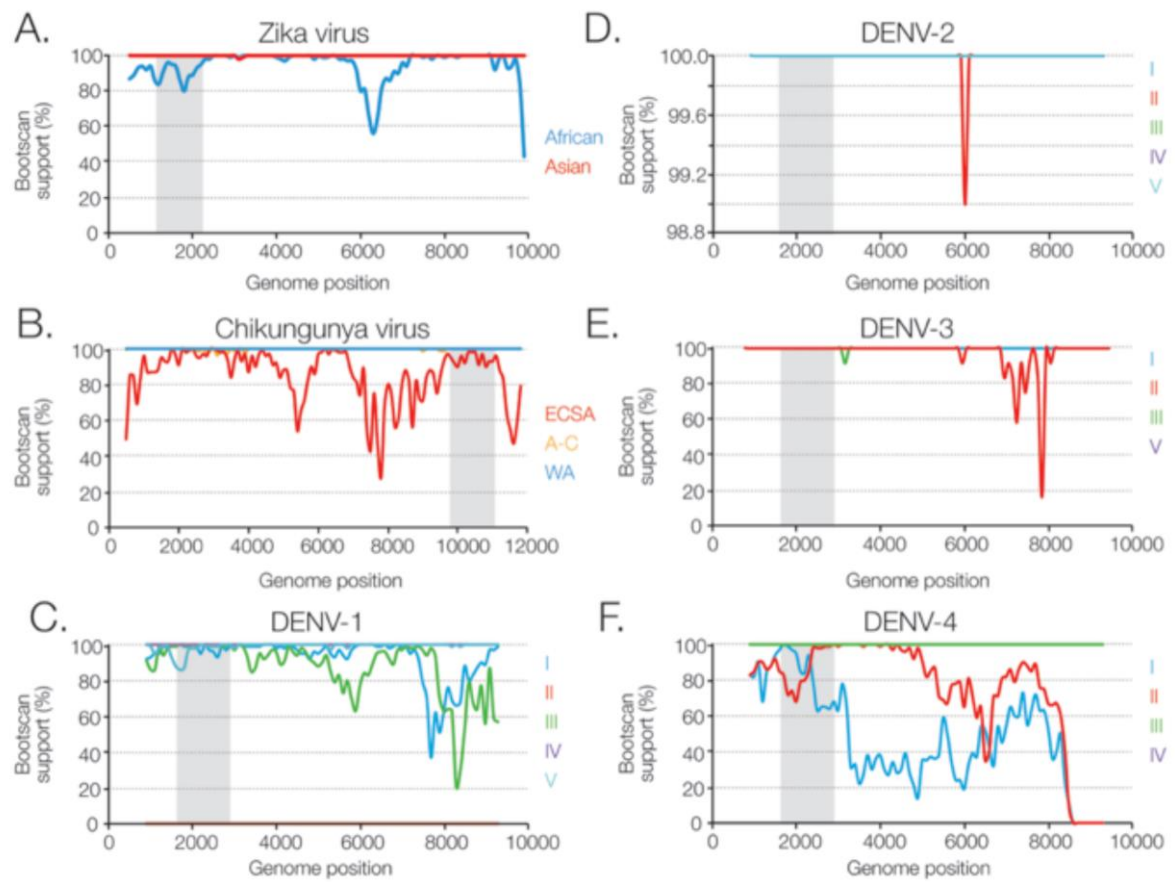
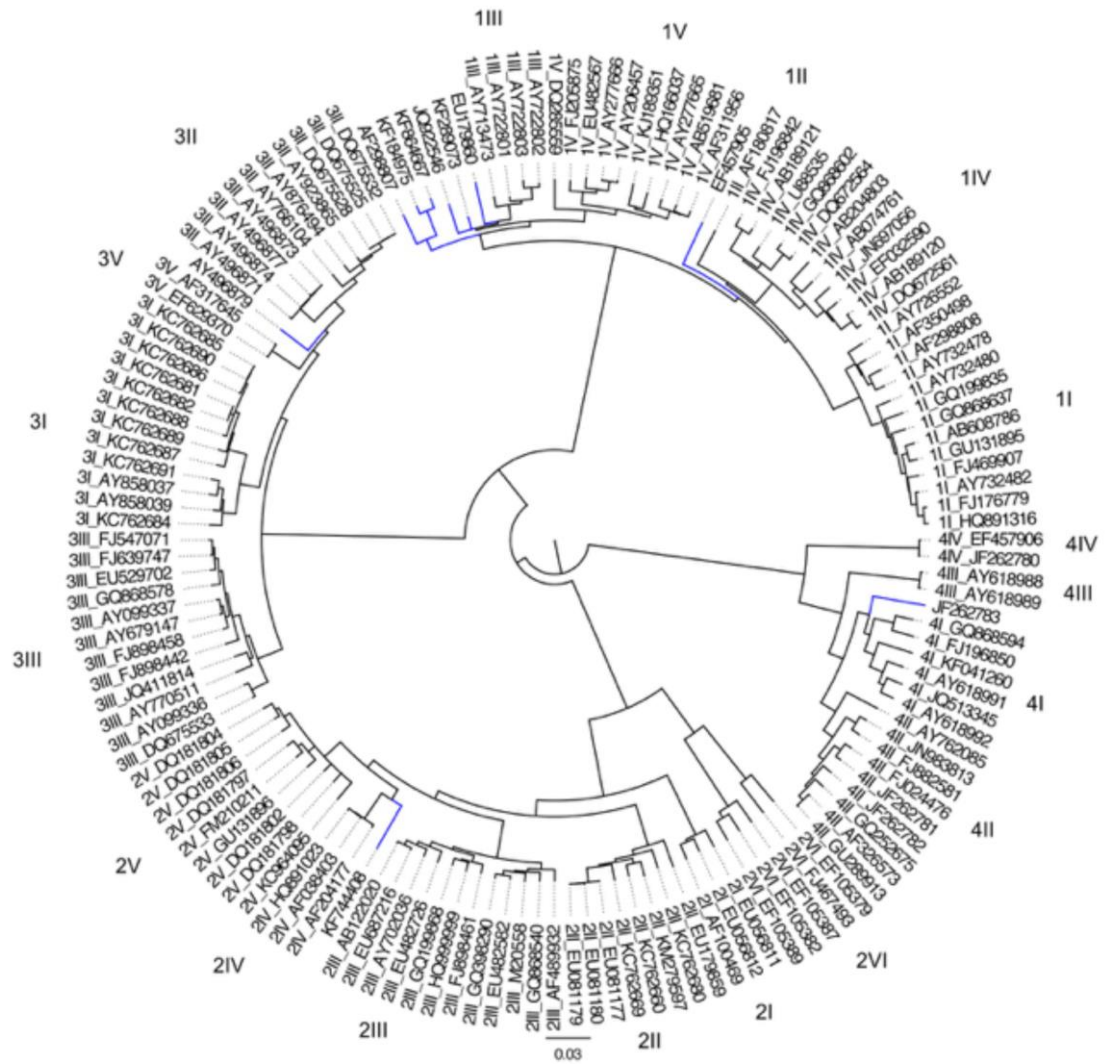
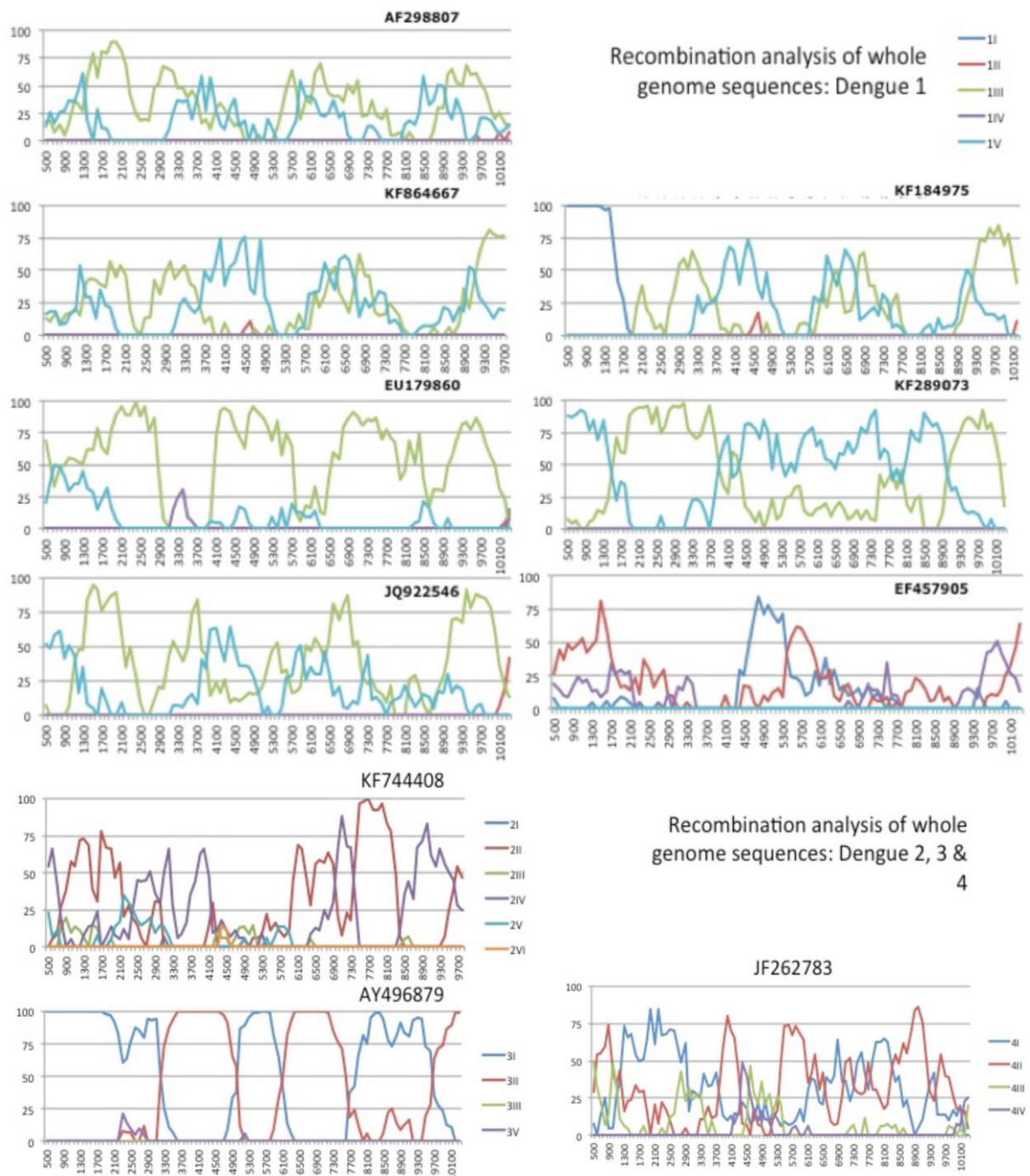


Figure 2: Graphical representation of the bootstrap support for reference sequences of ZIKV, CHIKV and DENV1-4 genotypes. The graphic was constructed using bootstrap results from NJ trees (1000 bootstrap replicates). The bootscanning method uses a sliding window of a 1,500bp segment that moves with steps of 100bp along the genome. The X-axis represents genome position, and the Y-axis represents bootstrap support. The grey lines mark the envelope gene.



Supplementary Figure 3: Maximum likelihood phylogenetic tree showing, in blue, the 10 whole genomes of DENV that could not be classified at genotype level, either manually or by our automated method. The 10 genomes were outliers falling between known Dengue genotypes.



Supplementary Figure 4: Bootscan results for the 10 whole genomes of DENV that could not be classified at genotype level (Supple Fig 2). Bootscanning analysis was performed by the typing tool, using a window length of 1,000bp and step size of 100bp, constructed the graphics. The different colours represent the genotypes for each serotype. The Y axis shows bootstrap results and the X-axis the position in the genome. In total, 7 sequences from DENV 1 were analysed and 1 sequence for each of the other serotypes, i.e. DENV2, DENV3 and DENV4.

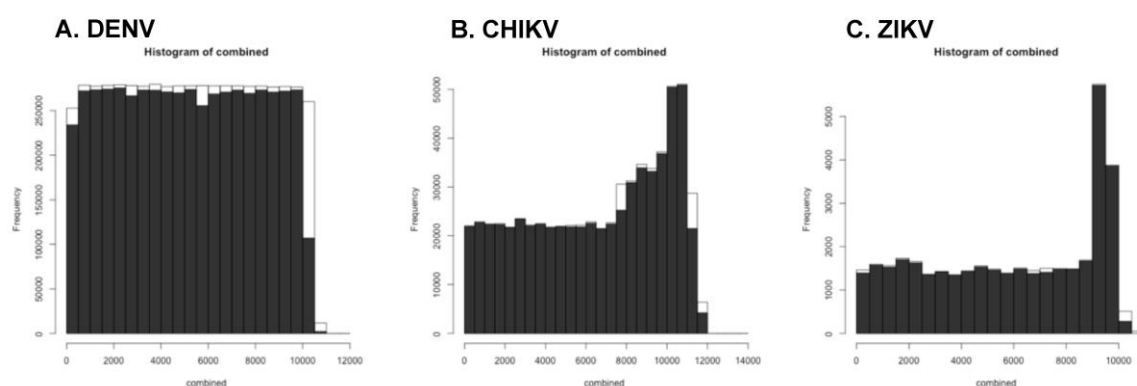
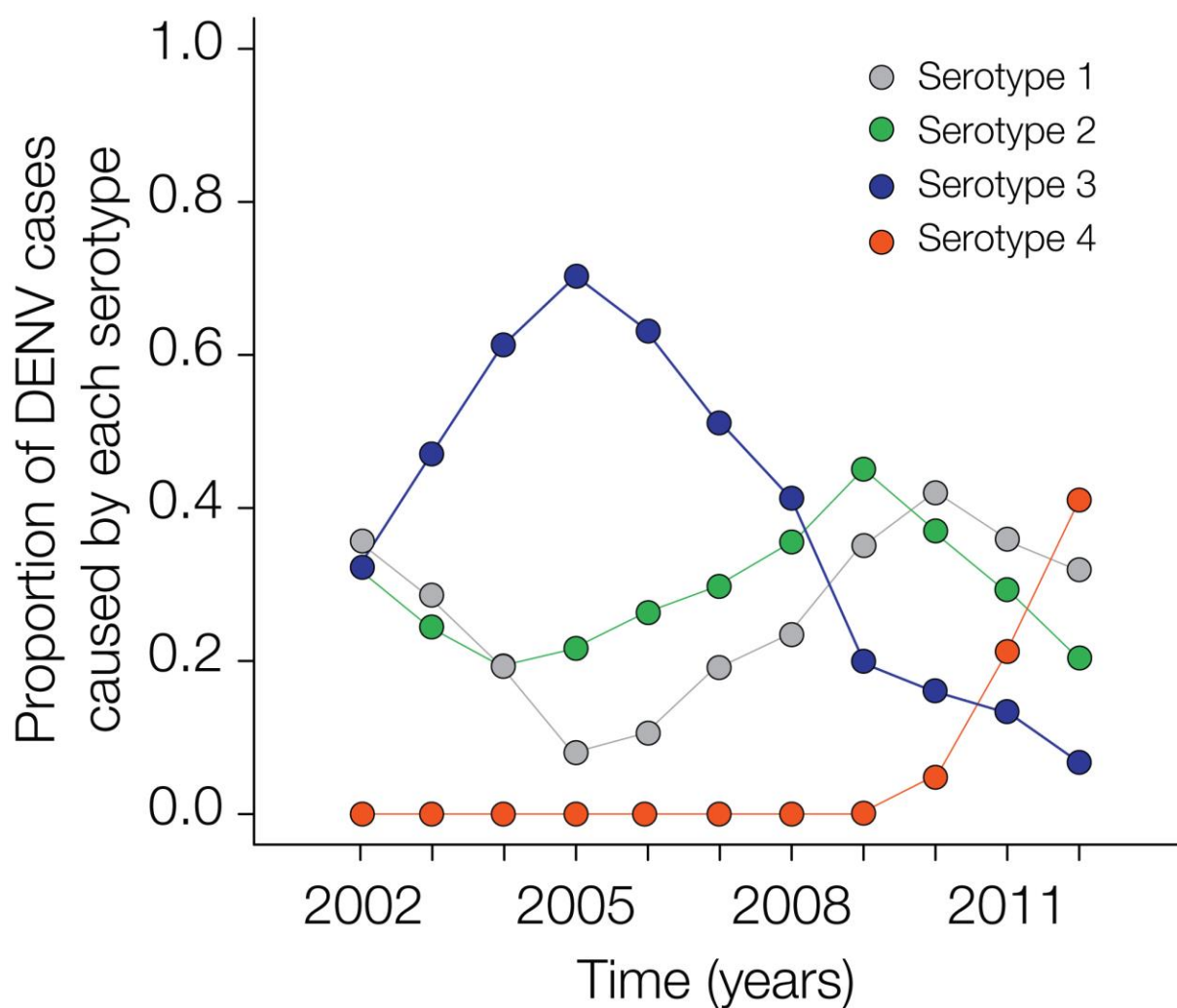


Figure 5. Histogram showing the distribution of the short reads (150bp) across the genome the genome. In black, reads correctly assigned. In white, reads unassigned. X-axis genome position, Y-axis, number of short reads.



Supplementary Figure 6: Proportion of DENV serotypes of all publicly available DENV1-4 sequences longer than 150nt collected in Brazil.

4.2 DESENVOLVIMENTO DAS FERRAMENTAS DE IDENTIFICAÇÃO VIRAL, E DA GENOTIPAGEM AUTOMÁTICA DOS DENV, CHIKV, ZIKV E YFV

Foram desenvolvidas 5 ferramentas, 4 de genotipagem automatizada e 1 de identificação viral. As ferramentas de genotipagem foram para o vírus dengue “*Dengue Virus Typing Tool*” (<http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/dengue>), para o vírus chikungunya “*Chikungunya Virus Typing Tool*” (<http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/chikungunya>), para o vírus zika “*Zika Virus Typing Tool*” (<http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/zika>) e a mais recente para o vírus da febre amarela “*Yellow Fever Virus Typing Tool*” (<http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/yellowfevervirus>). A ferramenta de identificação do tipo de vírus chamada “*Dengue, Zika, Chikungunya & Yellow Fever Viruses Typing Tool*” (<http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/aedesviruses>).

A ferramenta de identificação possui duas etapas e a de genotipagem possui quatro etapas. As duas primeiras etapas são idênticas para as 5 ferramentas e as duas etapas finais são exclusivas das ferramentas de genotipagem. Na primeira etapa, por meio de um formulário *web*, o usuário possui duas opções, na primeira é possível adicionar a(s) sequência(s) em um determinado campo ou fazer o *upload* do arquivo. Para as duas opções, a(s) sequência(s) deverão estar no formato *fasta* e clicar no botão “*start*” para analisar (Figura 9).



DENGUE, ZIKA, CHIKUNGUNYA & YELLOW FEVER VIRUSES TYPING TOOL

Dengue, Zika, Chikungunya & Yellow Fever Viruses Typing Tool Version 0.9 - Alpha

[Submit Job](#) [Monitor job](#) [How to cite](#) [Introduction to dengue, zika, chikungunya yellow fever classification](#) [How to use](#) [Example sequences](#)

Dengue, Zika, Chikungunya & Yellow Fever Viruses Typing Tool Version 0.9 - Alpha

This tool is designed to use Blast and phylogenetic methods in order to identify the Dengue, Zika, Chikungunya & Yellow Fever Viruses serotypes and genotypes of a nucleotide sequence.
Note for batch analysis: The tool accepts up to 2000 sequences at a time.

You may either:

- A. paste one or more sequences in FASTA format in the input field.
- B. upload a FASTA file.
- C. revisit results of a previous run

A) Paste nucleotide sequence(s) in FASTA format:

```
>T1_AF100469_Mexico_1992
TCCAGGCTTACCAATAATGGCCCAATCCTGGCATAACCATAGGAACGACCGATTCCAAAGAGTCC
TGATATTCATCCTACTGACAGCCATCGCTCCTCAATGACAATGCCTGCATAGGAATATCAAAATAGG
GACTTTGTGGAGAGTGTACAGAGGAGTGGGTGACATAGTTTGAACATGGAAGTTGTGTGAC
GACATGGCAAAATAACCAACACTGGCTTGAATGATAAAGACGAAACCAACACCCGCCA
CCTTAGGAAGTACTGTATAGAGCTAAACTGACCAACACAGACATCGCGCTGCCCAACACAA
GGGGAACCCACCTGAATGAAGAGCAGGACAAAGGTTTGTCTCAAACTTCTATGGTAGACAGAGG
ATGGGAAATGGATGTGGATTGTTTGGAAAGGAGGACATCGTACCTGTGTATGTTACATGCAAAA
AGAAATGTGAAGGAAATTTGTGACCCAAAGCTGGAAATGACTGTCTGATAGACCTCAATCA
GGGGAAGAACATGCACTGGAAATGACACAGGAACATGTAAGAAAGTCAAGATAACACCAAGAG
CTCCATCAGAGGCGGAAGTGAACAGCTATGGCACTGTTACGATGAGTGTCTTCCAAAGACGGCC
TCGACTTCAATGAGATGTGTGTGCAATGGAAGACAAAGCTTGGCTGGTGCACAGACAATGGTTC
>KR615989_Brazil_2015
ACTGGGCTCCACACTGGACAAACAAAGACACTGGTAGAGTTCAAGGACGACATGCCAAAGGCA
AACTGTGTGTTCTAGGAATCAAGAAGGACAGTTACACGCGCCTTGTGAGAGCTCTGGAGGCTG
```

B) Or, upload a FASTA with nucleotide sequences:

Selecionar arquivo... Nenhum arquivo selecionado.

C) Or, revisit results from a previous run:

Job-id:

Developed by: **FIOCRUZ/Bahia, Brazil** (Maria Inés Restovic, Marta Giovanetti, Wagner Fonseca, Murilo Freire, Luiz Alcantara), **KU Leuven, Belgium** (Kristof Theys, Pieter Libin, Lize Cuyper, Ana Abecasis, Anne-Mieke Vandamme), **Oxford, U.K.** (Nuno Faria and Oliver Pybus), **Evandro Chagas Institute, Brazil** (Marcio Roberto Teixeira Nunes), **CDC/OID/NCEZID** (Gilberto A. Santiago), **Emweb bvba, Belgium** (Koen Deforche) and **Africa Centre/UKZN, South Africa** (Tulio de Oliveira).

Contact: [Dr. Luiz Carlos Junior Alcantara](#), [Dr. Marcio Roberto Teixeira Nunes](#), [Dr. Nuno Faria](#) and/or [Prof. Tulio de Oliveira](#)










Figura 9. Página inicial das ferramentas de bioinformática.

Na segunda etapa, a(s) sequência(s) do usuário é submetida a análise do BLAST selecionando as sequências referências introduzidas na ferramenta (Tabela 1 do Anexo). Para que as sequências sejam genotipadas foi definido um corte maior igual 200pb para as ferramentas CHIKV, ZIKV, YFV, para DENV. Na terceira etapa as sequências são separadas individualmente e alinhadas com as referências utilizando o *software* ClustalW. Na quarta etapa é realizada a reconstrução de árvores filogenéticas no método NJ com valor 1.000 de *bootstrap*.

Para aumentar a velocidade e precisão dos resultados, foi adicionado a opção de *constraint backbone*, utilizando o modelo evolutivo HKY com gama e variação de taxas implementado no PAUP*. A classificação do genótipo da sequência é definida conforme o clado que a sequência se agrupa com valor de *bootstrap* igual ou superior a 70%. Caso o valor do *bootstrap* seja inferior a 70% a sequência é definida como “*unassigned*”. Para a ferramenta de identificação os resultados são realizados até a etapa 2.

A ferramenta produz um arquivo de saída no formato XML que contém todas as informações produzidas pelas análises. Esse arquivo XML é lido por um *script* e cria os relatórios resumidos desses resultados nos formatos HTML, XLS e CSV.

O relatório gerado pela ferramenta de identificação do vírus no formato HTML possui um gráfico no estilo pizza e uma tabela contendo as seguintes informações: tipo do vírus identificado, quantidade de sequências submetida, porcentagem da quantidade e uma legenda para se referenciar o gráfico (Figura 10). No tipo do vírus é gerado um *link* do nome que, ao ser clicado, o usuário será redirecionado para uma das quatro ferramentas de genotipagem criadas para ter informações mais detalhadas sobre determinada sequência. Nos relatórios no formato XLS e CSV é possível encontrar o nome da sequência, o tipo identificado, tamanho, início e final na região genômica do vírus.

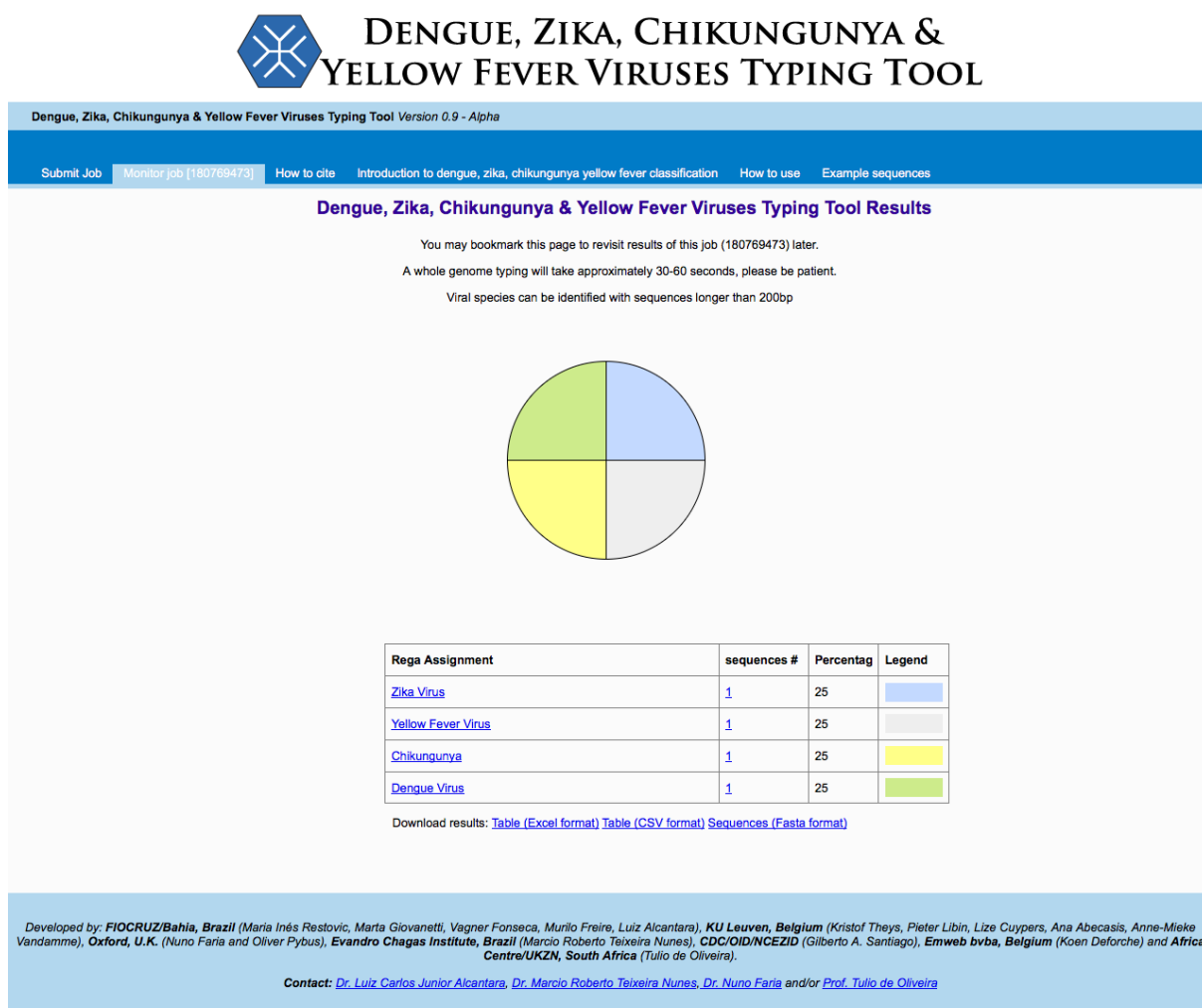


Figura 10. Relatório em HTML da ferramenta de identificação dos patógenos contendo informações sobre tipo do vírus, quantidade de sequências, porcentagem da quantidade e uma imagem da legenda do gráfico.

Os resultados das ferramentas de genotipagem no formato XLS e CSV contém um resumo de todos os resultados, incluindo o nome da sequência, comprimento, posições inicial e final sobre o genoma, o vírus atribuído e o genótipo. O relatório HTML contém informações sobre o nome da sequência, comprimento da sequência, tipo do vírus, o genótipo, um link para um relatório detalhado e uma ilustração do genoma do vírus preenchida a região genômica no qual sequência faz parte, sendo possível fazer o *download* dos relatórios nos formatos XLS, CSV, XML e as sequências submetidas no formato fasta (Figura 11).

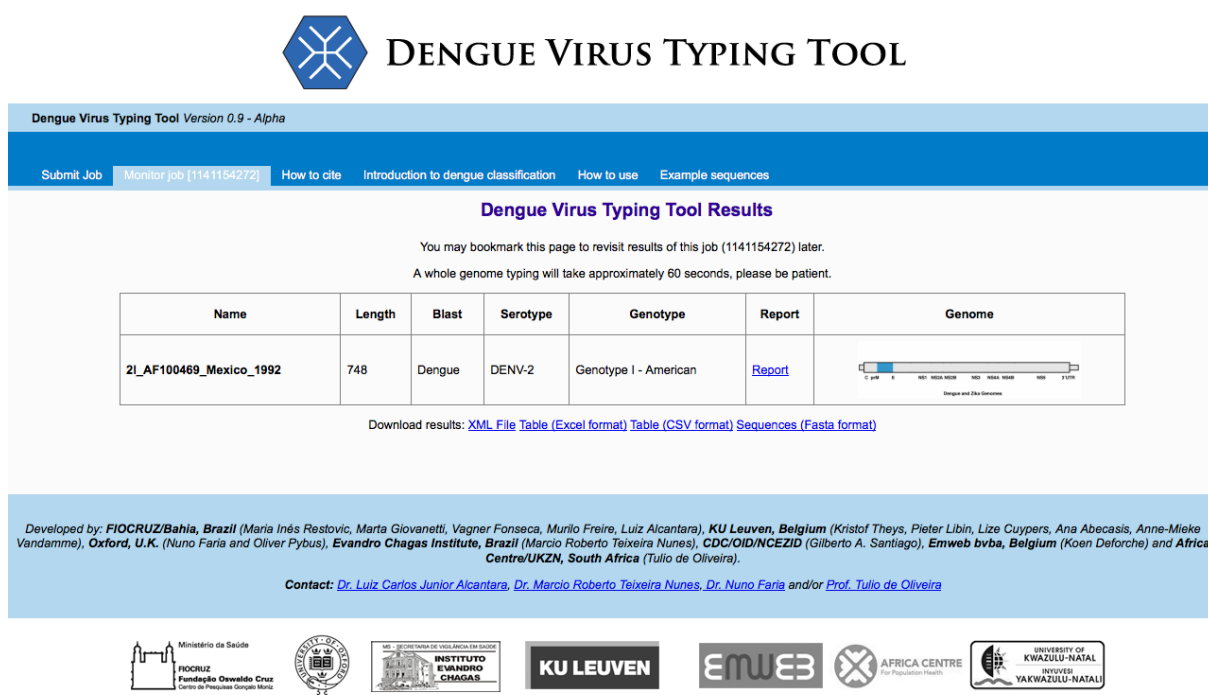


Figura 11. Relatório em HTML da ferramenta de genotipagem contendo informações sobre o nome da sequência, comprimento, espécies virais atribuídas, genótipo, e uma ilustração do genoma do vírus.

O *link* relatório gerado individualmente para cada sequência analisada é composto por duas sessões que são chamadas de “*Sequence Assignment*” e “*Phylogenetic Analysis Details*”. Na primeira sessão, são apresentadas informações sobre a sequência como: tipo do vírus, genótipo do vírus com o valor de *bootstrap* da análise da sequência, uma figura ilustrativa com região genômica preenchida com a posição inicial e final da sequência. Na segunda opção são apresentadas opções para *download* do alinhamento no formato fasta ou NEXUS, a árvore filogenética criada nos formatos PDF e NEXUS. Uma imagem dessa árvore é plotada com fácil interpretação pois a sequência analisada é colocada no topo da árvore e um log da análise

realizada pelo PAUP* dessa sequência, que contem o modelo evolutivo e os parâmetros utilizados (Figura 12).

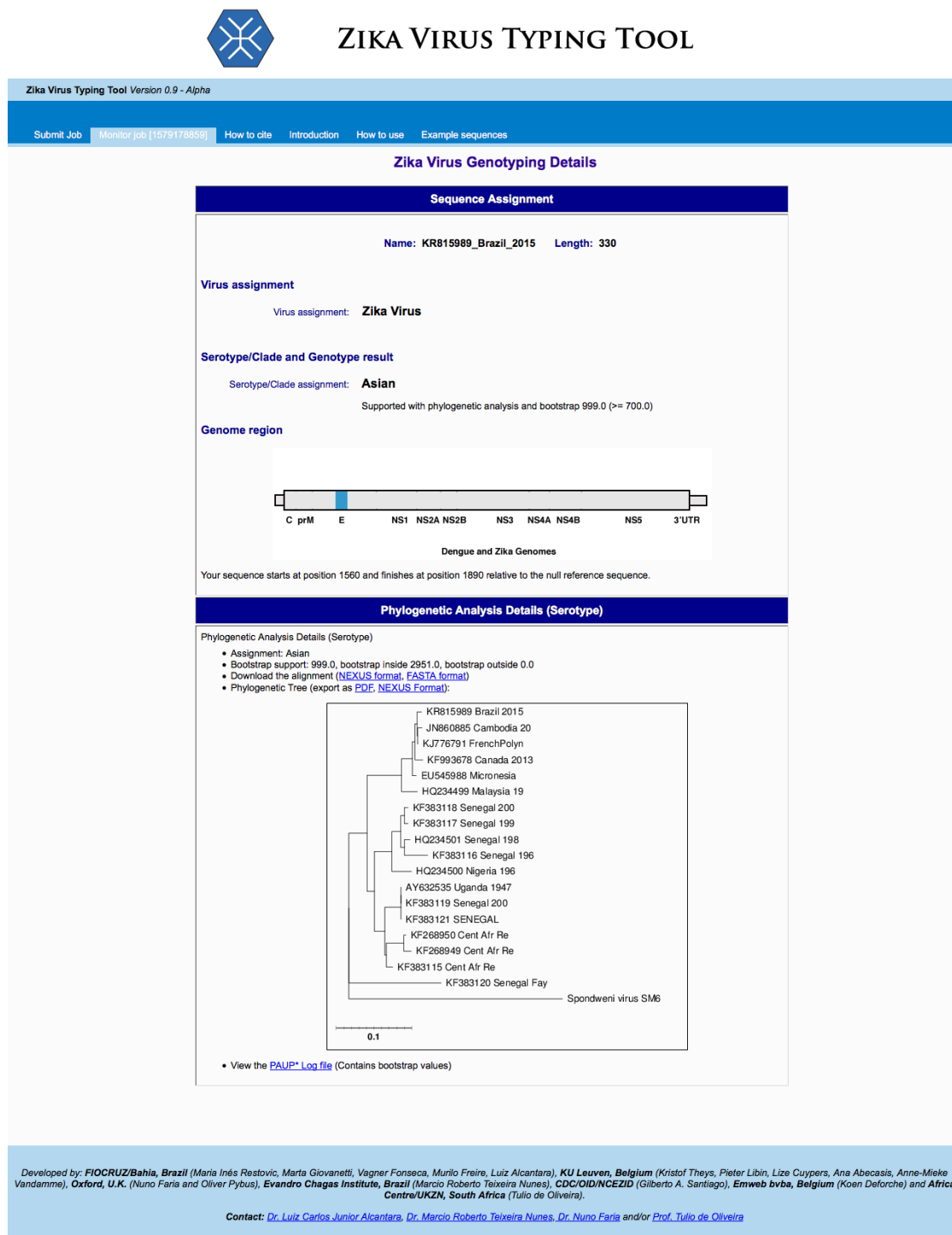


Figura 12. O relatório detalhado em HTML que contém informações como: o nome da sequência, comprimento, vírus atribuído, genótipo, ilustração do genoma viral, análise filogenética, alinhamento e árvore filogenética reconstruída.

4.3 ANÁLISE DO DESEMPENHO DA FERRAMENTA DO YFV

O processo de avaliação envolveu a classificação de todos os genomas disponíveis publicados completos e regiões do envelope do YFV. Foi analisado um total de 137 YFV. Em vez de selecionar um número fixo de sequências (por exemplo, 10 ou 100 por genótipo), foram analisados todos os dados publicados que continham a região do envelope para descobrir se o nosso método automatizado poderia classificar todas as sequências disponíveis. Cada conjunto de dados das sequências coletada foram analisadas utilizando os métodos ML, e análise Bayesian mencionada anteriormente e o método automatizado feito pela ferramenta de genotipagem. Para terminar a precisão da ferramenta de genotipagem, calculou-se a sensibilidade com a fórmula $TP / TP + FN$ e a especificidade com a fórmula $TN / TN + FP$, onde TP = verdadeiro positivo, FP = falso positivo, TN = verdadeiro negativo e FN = falso negativo (BANO et al., 2010). Os resultados foram submetidos pelo algoritmo desenvolvido e acoplado ao *framework* Rega-Genotype para realização dos cálculos de sensibilidade, especificidade.

Os resultados para genoma completo são demonstrado na tabela 1, a sensibilidade e especificidade variou de 0% a 100%. Para YFV, foi obtido 100% de especificidade e sensibilidade na região envelope do vírus.

Tabela 1. Análise filogenética de genomas completos realizado ferramenta de genotipagem para classificar o YFV.

Espécie de Vírus	Conhecidos	TP	TN	FP	FN	SENS	SPEC	ACC
Vírus Febre Amarela								
Sul Americano I	47	47	3399	0	0	100%	100%	100%
Sul Americano II	38	38	4234	0	0	100%	100%	100%
Oeste Africano	40	40	4128	0	0	100%	100%	100%
Leste Africano	12	12	4072	0	0	100%	100%	100%
Total de sequências	137							

Os resultados da classificação foram comparados com análise filogenética manual. Em TP = verdadeiro positivo, TN = verdadeiro negativo, FP = falso positivo, FN = falso negativo, SENS = sensibilidade, SPEC = especificidade.

5 DISCUSSÃO

A mineração automática das sequências coletadas no GenBank foi importante para selecionar somente aquelas de interesse para construir o conjunto de sequências referências utilizadas pelas ferramentas. A utilização de um Banco de Dados Temporário, fornecido por um Sistema de Gerenciamento de Banco de Dados (SBDG), foi fundamental na organização, armazenamento e atualização das sequências e de suas respectivas informações. Um banco de dados relacional permitiu facilidades nos acessos aos dados e em consultas específicas. Possibilitando selecionar com mais facilidades as sequências que foram utilizadas no processo de construção das sequências referências e na seleção das sequências que foram utilizadas para o teste de eficiência da ferramenta automatizada.

A coleta das informações de cada sequência foram coletadas inicialmente no *Genbank*, tendo-se o cuidado em posteriormente checar a veracidade dessas anotações, pois segundo Guimarães (2006), um dos problemas mais comuns relatados no processo de coleta de informações de banco de dados público é o erro e/ou a falta de padronização nos parâmetros das sequências armazenadas. Para checar a veracidade e colher informações faltantes no banco de dados públicos, houve a necessidade de consultar os artigos publicados, para descobrir alguns sorotipos (DENV), genótipos e região geográfica (DENV, CHIKV, ZIKV e YFV) de algumas sequências.

Considerando as dificuldades encontradas na aquisição dessas informações, ressalva-se a importância do fornecimento e disponibilidade pública dos dados gerados, respeitando a propriedade intelectual, quando estes forem factíveis para o conhecimento científico sobre os arbovírus estudados nessa dissertação.

As ferramentas automatizadas de bioinformáticas desenvolvidas nesta dissertação proporcionam a comunidade mundial a classificação precisa para diferentes arbovírus. A ferramenta proporciona o alinhamento e a reconstrução de árvores filogenéticas para cada segmento de sequência introduzida na ferramenta com as sequências referências utilizadas pela ferramenta, sendo possível realizar análise *bootscanning* e tirar conclusões estatisticamente para um determinado genótipo. Foi introduzido um corte de maior igual a 70% para as análises de *bootstrap* e maior igual a 90% para as análises do *bootscan* reduzindo assim riscos nas classificações minimizando assim falso positivos aumentando assim a qualidade nos resultados fornecidos pela ferramenta.

Foi priorizado nesse projeto uma interface com usabilidade facilitada, para que os usuários tenham um uso simplificado, considerando-se o acesso de usuários mais ou menos experientes, e que a sua utilização seja eficaz, produtiva, além de segura. Foram adicionadas novas funcionalidades a interface no *framework* como a possibilidade de consultar trabalhos processados anteriormente, facilitando assim uma busca por resultados futuros, além de adicionados *scripts* para processar os resultados fornecidos pela ferramenta podendo assim calcular a sensibilidade e especificidade dos resultados.

6 CONCLUSÕES

Em conclusão, o desenvolvimento deste método computacional permite a classificação com alto rendimento para genótipos das espécies de DENV, ZIKV, CHIKV e YFV. As espécies podem ser classificadas a partir de uma leitura curta (*short reads*) de qualquer arquivo oriundo de plataformas Sequenciamento de Nova Geração (NGS) (*Next-Generation Sequencing*), tais como metagenômica do RNA-seq *Illumina*. Os genótipos são classificados com mais confiança usando a região envelope ou sequências completas do vírus. Os *softwares* estão disponíveis para acesso online gratuito hospedado nos servidores dedicados do BioAfrica (<http://www.bioafrica.net/>), na opção do menu *softwares* ou por endereço completo descrito no item 6 desta dissertação.

TRABALHOS FUTUROS

Nessa seção estão listados alguns pontos que foram identificados como oportunidades de evolução do trabalho realizado nessa dissertação. Alguns dos tópicos listados têm como objetivo tornar o *framework* do Rega-Genotype mais completo, explorando aspectos que auxiliaram pesquisadores futuros na adição de ferramentas de genotipagem para vírus.

- O desenvolvimento de classes genéricas para serem adicionadas ao *framework* Rega-Genotype, ao usuário informar um conjunto de sequências de um determinado vírus, quais serão mais representativas na construção das sequências referências, demonstrando, também, qual a melhor região do genótipo do vírus será possível realizar o processo automático nas reconstruções de árvore filogenética.
- O desenvolvimento de classes genéricas para serem adicionadas ao *framework* Rega-Genotype para genotipar com eficiência arquivos no formato FASTQ proveniente de *softwares* NGS, realizando, também a genotipagem dos vírus das ferramentas que foram criadas usando o Rega-Genotype

REFERÊNCIAS

- ADDRIANS, P.; ZANTINGE, D. **Data Mining**. Inglaterra: Addison-Wesley, 1996.
- ALCANTARA, L. C. et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. **Nucleic Acids Research**, v. 37, p. W634–W642, 2009.
- BARRETT, A. D.; HIGGS, S. Yellow fever: a disease that has yet to be conquered. **Annual Rev. Entomol.**, v. 52, p. 209-229, 2007.
- BIOINFORMATICS FACTSHEET. Disponível em: <http://www.uio.no/studier/emner/matnat/ifi/INF4350/h07/undervisningsmateriale/Bioinformatics%20and%20Molecular%20Genetics%20_%20Intro%20NCBI.pdf>. Acesso em: 1 jul. 2016.
- BRACHNAD, R.J. e ANAND, T. **The process of knowledge discovery in databases**. In: FAYYAD, U. M. et al. **Advances in knowledge discovery in data mining**. Palo Alto, CA: AAAI, 1996.
- BRAY, RS. **Armies of Pestilence: The impact of disease on history**. James Clarke & Co., 2004. 258p.
- BRECK, S. In Albert Bushnell Hart, ed. **American History Told by Contemporaries**. New York: Macmillan, 1929. v. 3
- BRYANT, JE; *et al.* **Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas**. PLoS Pathog, 2007, 3(5), e75.
- CARDOSO, C. W. et al. Outbreak of exanthematous illness associated with zika, chikungunya, and dengue viruses, Salvador, Brazil. **Emerg. Infect. Dis.**, v. 21, n. 12, p. 2274-2276, 2015.
- CASSADO, S; *et al.* **Emergence of chikungunya fever on the French side of Saint Martin Island, October to December 2013**. Euro Surveill. 2014; 19, n.13, p.1-4.
- CDC. Centers for Disease Control and Prevention. **Yellow Fever**. Disponível em <http://www.cdc.gov/yellowfever/>. Acesso em: 01 set. 2016.
- CHAMBERS, T. J. et al. Flavivirus genome organization, expression, and replication. **Annual Rev. Microbiol.**, v. 44, p. 649-688, 1990.
- CHEVENET, F. et al. Searching for virus phylotypes. **Bioinformatics**, v. 29, n. 5, p. 561–570, 2013.

CLETON, N. et al. Come fly with me: review of clinically important arboviruses for global travelers. **J. Clin. Virol.**, v. 55, n. 3, p. 191-203, 2012.

COLEMAN, W. Epidemiological method in the 1860s: yellow fever at Saint-Nazaire. **Bull. Hist. Med.**, v. 58, n. 2, p. 145–163, 1983,

COOK, S.; HOLMES, E. C. A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: Flaviviridae) and the evolution of vector transmission. **Arch. Virol.**, v. 151, p. 309-325, 2006.

COOK, S. et al. Molecular evolution of the insect-specific flaviviruses. **J. Gen. Virol.**, v. 93, p. pt2, p. 223-234, 2012.

_____. Isolation of a novel species of flavivirus and a new strain of Culex flavivirus (Flaviviridae) from a natural mosquito population in Uganda. **J. Gen. Virol.**, v. 90, p. pt11, p. 2669-2678, 2009.

DE OLIVEIRA, T. et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. **Bioinformatics**, v. 21, n. 19, p. 3797-800, 2005.

DE SOUZA, R. P. et al. Detection of a new yellow fever virus lineage within the South American genotype I in Brazil. **J. Med. Virol.**, v. 82, p. 175–185, 2010.

DEUBEL, V. et al. Dengue 2 virus envelope protein expressed by a recombinant vaccinia virus fails to protect monkeys against dengue. **J. Gen. Virol.**, v. 69, n. Pt 8, p. 1921-1029, 1988.

DICK, G.W.; KITCHEN, S.F.; HADDOW, A.J. Zika virus. I. Isolations and serological specificity. **Trans. R. Soc. Trop. Med. Hyg.**, v. 46, p. 509-520, 1952.

DILLY, R. **Data Mining: an introduction**. Belfast: Parallel Computer Centre, Queens University, 1999.

DINIZ, C.A. e LOUZADA-NETO, F. **Data Mining: uma introdução**. São Carlos: Associação Brasileira de Estatística, 2000.

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

ELMASRI, R.E.; NAVATHE, S. **Sistemas de banco de dados**. 4. ed. Addison-Wesley, 2005.

FAYE, O. et al. Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. **PLoS Negl. Trop. Dis.**, v. 8, n. 1, p. e2636, 2014.

FAYYAD, U.M. et al. **The KDD process for extracting useful knowledge from volumes of data.** In: _____. Advances in knowledge discovery in data mining. menlo park: AAAI Press, 1996.

FIGUEIREDO, L.T.M. Emergent arboviruses in Brazil. **Rev. Soc. Bras. Med. Trop.**, v. 40, p. 224-229, 2007.

FLAMAND, M. et al. Dengue virus type 1 nonstructural glycoprotein ns1 is secreted from mammalian cells as a soluble hexamer in a glycosylation-dependent fashion. **J. Virol.**, v. 73, n. 7, p. 6104-6110, 1999.

GOLLINS, S.W.; PORTERFIELD, J. S. Flavivirus infection enhancement in macrophages: an electron microscopic study of viral cellular entry. **J. Gen. Virol.**, v. 66, n. pt9, p. 1969-1982, 1985.

GOULD, E. A. et al. Origins, evolution, and vector/host coadaptations within the genus Flavivirus. **Adv. Virus Res.**, v. 59, p. 277–314, 2003.

GUBIO, S. et al. **Zika Virus Outbreak, Bahia, Brazil. Emerg. Infect. Dis.**, v. 21, n. 10, 2015.

GUBLER, D. J.; CLARK, G.G. Dengue/dengue hemorrhagic fever: the emergence of a global health problem. **Emerg. Infect. Dis.**, v. 1, n. 2, p.55-57, 1995.

GUBLER, D.J. Dengue and Dengue Hemorrhagic fever. **Clin. Microbiol. Rev.**, v. 11, n. 3, p. 480–496, 1998.

_____. Human arbovirus infections worldwide. **Ann. N. Y. Acad. Sci.**, v. 951, n. 1, p. 13-24, 2001.

_____. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. **Trends Microbiol.**, v. 10, n. 2, p. 100-103, 2002.

_____. et al. Flaviviruses. In: KNIPE, D.M.; HOWLEY, P.M. (eds.) **Fields Virology**, 5th ed. Philadelphia, PA: Wolters Kluwer and Lippincott Williams & Wilkins, 2007. p.1153–1252.

GUJARATI, D.N. **Econometria Básica.** Trad. Ernesto Yoshita. São Paulo: Makron Books, 2000.

HADDOW, A. J. X.-The Natural History of Yellow Fever in Africa. **Proc. Roy. Soc. Edinburgh. Sec. B. Biology**, v. 70, n. 3, p. 191–227, 2012.

HALSTEAD, S.B. **Dengue vaccine development: a 75% solution?** The Lancet.com, Bethesda, 2012.

HAMMON, W.M.C.D. et al. New hemorrhagic fevers of children in the Philipines and Thailand. **Trans. Ass. Am. Phys.**, v.73, p.140-155, 1960.

HAND, D.J. Data Mining: statistics and more? **The Am. Statist.**, v. 52, n. 2, p. 112-118, 1998.

HARRIS, E. et al. **Molecular biology of flaviviruses**. Novartis Foundation Symposia, 2006. 277: 23-39.

HEINZ, F.X. e STIASNY, K. Flaviviruses and flavivirus vaccines. **Vaccine**, v. 30, n. 29, p. 4301-4306, 2012.

HOLMES, E.C. TWIDDY, S.S. The origin, emergence and evolutionary genetics of dengue virus. **Infect. Genet. Evol.**, v. 3, p. 19-28, 2003.

HORWOOD, P; et al. The threat of chikungunya in Oceania. West. Pac. Surveill. **Response J.**, v. 4, p. 8-10, 2013.

HOSHINO, K; et al. Isolation and characterization of a new insect flavivirus from Aedes albopictus and Aedes flavopictus mosquitoes in Japan. **Virology**, v. 391 n. 1, p. 119-129, 2009.

JARCHO, S. John Mitchell, Benjamin Rush, and yellow fever. **Bull Hist. Med.**, v. 31 n. 2, p. 132-6, 1957.

JAWETZ, E. et al. **Microbiologia médica**. 22 rd ed. Rio de Janeiro: Mc Graw Hill; p. 411-26, 2005.

KIMURA, R; HOTTA, S. Studies on dengue virus (VI). On the inoculation of dengue virus into mice. **Nippon Igaku Hoshasen Gakkai Zasshi**, v. 3379, p. 629-633, 1944.

KRAEMER, M.U. et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. **Elife**, 2015.

KULARATNE, S.A. Dengue fever. **BMJ**, v. h4661. p. 351, 2015

LEMOS, M. **Workflow para Bioinformática**. Thesis(Doutorado)- PUC. Departamento de Informática. Rio, 2004.

LEVINE, D.M. et al. **Estatística: teoria e aplicações**. Trad. Teresa C.P. de Souza. Rio de Janeiro: LTC Editora, 2000.

LINDENBACH, B.D.; RICE, C. M. Flaviviridae: the viruses and their replication. In KNIPE, D.M. AND HOWLEY, P.M. (Eds.) **Fields Virology**. Lippincott-Williams & Wilkins, Philadelphia, PA, 2001. p. 991-1041.

LINDENBACH, B.D.; RICE, C. M. Molecular Biology of Flavivirus. In: CHAMBERS, T.J.; MONATH, T.P. **Advances in Virus Research: The Flavivirus:structure, replication and evolution**. California, Academic Press, 2003. v. 59, p. 23-61.

LINDENBACH, B.D; et al. Flaviviridae: The Viruses and Their Replication. In: KNIPE, D. M.; P. M. H. **Fields Virology**. Philadelphia, PA. 5 ed. Lippincott Williams & Wilkins. 2007. p. 1101.

LIU, W. J. et al. Analysis of adaptive mutations in Kunjin virus replicon RNA reveals a novel role for the flavivirus nonstructural protein NS2A in inhibition of beta interferon promoter-driven transcription. **J. Virol.**, v. 78, n. 22, p. 12225-12235, 2004.

LO PRESTI, A. et al. Origin, evolution, and phylogeography of recent epidemic CHIKV strains. **Infect. Genet. Evol.** v. 12, n. 2, p. 392, 2012.

MA, L. et al. Solution structure of dengue virus capsid protein reveals another fold. **Proc. Nat. Acad. Scien.**, v. 101, n. 10, p. 3414-3419, 2004.

MACKOW, E. et al. The nucleotide sequence of dengue type 4 virus: analysis of genes coding for nonstructural proteins. **Virology**, v. 159, n. 2, p. 217-228, 1989.

MACNAMARA, F. N. Zika virus: a report on three cases of human infection during an epidemic of jaundice in Nigeria. **Trans. R. Soc. Trop. Med. Hyg.**, v. 48, p. 139-145, 1954.

MAIHURU, A.T.A. et al. Dengue: an arthropod-borne disease of global importance. **European Journal Clinical Microbiology & Infectious Diseases, Heidelberg**, v. 23, p. 425-433, 2004.

MANDL, CW; et al. Genome sequence of tick-borne encephalitis virus (Western subtype) and comparative analysis of nonstructural proteins with other flaviviruses. **Virology**, v. 173 n. 1, p. 291-301, 1989

MANNILA, H. Data mining: machine learning, statistics and databases. In: INTERNATIONAL CONFERENCE ON STATISTICS AND SCIENTIFIC DATABASE MANAGEMENT, Estocolmo, 1996. v. 8.

MARCHETTE, N.J. et al. Isolation of Zika virus from Aedes aegypti mosquitoes in Malaysia. **Am. J. Trop. Med. Hyg.**, v. 18, p. 411-415, 1969.

MARON, D.F. **First dengue fever vaccine gets green light in 3 countries**. Scientific American. Retrieved 3 Feb 2016. Disponível em <https://www.scientificamerican.com/article/first-dengue-fever-vaccine-gets-green-light-in-3-countries/>. Acesso em: 29 ago. 2016.

MARTINS, G.A. **Estatística Geral e Aplicada**. São Paulo: Atlas, 2001.

MATTAR, F.N. **Pesquisa de Marketing**. São Paulo: Atlas, 1998.

MCNEILL, J.R. **Mosquito Empires: Ecology and War in the Greater Caribbean, 1620–1914**. 1ed., Cambridge University Press, 2010. p. 390,

_____. Yellow Jack and Geopolitics: Environment, Epidemics, and the Struggles for Empire in the American Tropics, 1650-1825. **OAH Magazine of History**, v. 18, n. 3, p. 9–13, 2004.

MEERS, P.D. Yellow fever in Swansea, 1865. **J. Hyg.**, v. 97, n. 1, p. 185–191, 1986.

MODIS, Y. et al. Structure of the dengue virus envelope protein after membrane fusion. **Nature**, v. 427, n. 6972, p. 313-319, 2004.

MORETTIN, P.A.; TOLOI, C.M. **Séries Temporais**. 2 ed. São Paulo: Atual, 1987.

MOUREAU, G. et al. Flavivirus RNA in Phlebotomine Sandflies. **Vector Borne Zoonotic Diseases**. v. 10, n. 2, p. 195-197, 2010.

MUKHOPADHYAY, S. et al. A structural perspective of the Flavivirus life cycle. **Nature Reviews Microbiology**, v. 3, n. 1, p. 13-22, 2005.

MUSSO, D. et al. Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. **Euro Surveill**, v. 19, 2014.

OLSON, J.G. et al. Zika virus, a cause of fever in Central Java, Indonesia. **Trans. R. Soc. Trop. Med. Hyg.**, v. 75, p. 389-393, 1981.

PADOVANI, C.R. **Estatística na Metodologia da Investigação Científica**. Botucatu: UNESP, 1995.

PAN AMERICAN HEALTH ORGANIZATION Number of reported cases of Chikungunya Fever in the Americas, by country or territory 2013-2015. **Epidemiological Week / EW 24**, 2015.

PEREIRA, J.C.R. **Análise de Dados Qualitativos**. São Paulo: Edusp/Fapesp, 1999.

PERERA, R.; KUHN, R.J. Structural proteomics of dengue virus. **Curr Opin Microbiol.**, v. 11, n. 4, p. 369-377, 2008.

PIERSON, T.C.; DIAMOND, M.S. Flavivirus. In: FIELDS, B.N.; KNIPE, D.A.M.; HOWLEY, P.M. (Eds.) **Fields Virology**. 6 ed., 2013. p. 747-794.

POWELL, J. H. **Bring Out Your Dead:** The great plague of yellow fever in Philadelphia in 1793. University of Pennsylvania Press, 1949.

POWERS, A.M.; LOGUE, C.H. Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. **J. Gen. Virol.** v. 88, p. 2363–2377, 2007.

PROSDOCIMI, F. et al. Bioinformática: Manual do Usuário. **Biotecnologia Ciência & Desenvolvimento**, n. 29, 2002.

REY, F.A. et al. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. **Nature**, v. 375 n. 6529, p. 291-298, 1995.

RIGAU-PEREZ, J.G. et al. Dengue and dengue haemorrhagic fever. **Lancet**, v. 352, p. 971-977, 1998.

ROBINSON, M.C. An epidemic of virus disease in southern province, Tanganyika territory, in 1952-53. **Trans. R. Soc. Trop. Med. Hyg.**, v. 49, n. 2832, 1955.

RODRIGUES, F.N. et al. Epidemiology of Chikungunya Virus in Bahia, Brazil, 2014-2015. **PLOS Currents Outbreaks**. 1 ed., 2016.

ROSS, R.W. The Newala epidemic III. The virus: isolation, pathogenic properties and relationship to the epidemic. **J. Hyg.** v. 54, n. 17791, 1956.

ROTH, A; et al. Concurrent outbreaks of Dengue, Chikungunya and Zika virus infections - an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012-2014. **Euro Surveill.**, v. 41, p. 16-19, 2014.

SABIN, A.B.; SCHLESINGER, M.C. Production of immunity to dengue with virus modified by propagation in mice. **Science**, v. 101, p. 640-642, 1945.

SADE, A.S.; SOUZA, J.M. **Prospecção de Conhecimento em Bases de Dados Ambientais**. Rio de Janeiro: UFRJ, 1996.

SÁNCHEZ-SECO, M.P; et al. Generic RT-nested-PCR for detection of flaviviruses using degenerated primers and internal control followed by sequencing for specific identification. **J. Virol. Meth.**, v. 126 n. 1-2, p. 101-109, 2005.

SANG, R.C. et al. Isolation of a new flavivirus related to cell fusing agent virus (CFAV) from field-collected flood-water Aedes mosquitoes sampled from a dambo in central Kenya. **Arch. Virol.**, v. 148 n. 6, p. 1085-1093, 2003.

SFAKIANOS, J. et al. **West Nile virus**. Foreword by David Heymann, 2 ed. New York: Chelsea House. 2009. p. 17.

SILVA, A.M. **Caracterização molecular dos vírus dengue circulantes em Pernambuco: implicações epidemiológicas**. 2013. Tese (Doutorado em Saúde Pública) - Fundação Oswaldo Cruz, Centro de Pesquisas Aggeu Magalhães, Recife, 2013.

SOLER, J.C. et al. A mortality study of the last outbreak of yellow fever in Barcelona City (Spain) in 1870. **Gaceta Sanitaria / S.E.S.P.A.S.** v. 23, n. 4, p. 295-299, 2008.

TOLLE, M.A. Mosquito-borne diseases. **Curr. Probl. Pediatr. Adolesc. Health Care**, v. 39, n. 4, p. 97-140, 2008.

VOLK, S.M. et al. Genome scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. **J. Virol.**, v. 84, p. 6497-6504, 2010.

VON LINDERN, J.J. et al. Genome analysis and phylogenetic relationships between east, central and west African isolates of Yellow fever virus. **J. Gen. Virol.**, v. 87, p. 895–907, 2006.

WANG, W.K. et al. Detection of dengue virus replication in peripheral blood mononuclear cells from dengue virus type 2-infected patients by a reverse transcription-real-time PCR assay. **J. Clin. Microbiol.**, v. 40, n. 12, p. 4472-4478, 2002.

WEAVER, S.C.; VASILAKIS, N. Molecular evolution of dengue viruses: Contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. **Infect. Genet. Evol.**, v. 9, p. 523-540, 2009.

WEISS, V.A. **Estratégias de Finalização da Montagem do Genoma da Bactéria Diazotrófica Endofítica *Herbaspirillum seropedicae* SmR1**. 72 f. Dissertação (Mestrado em Ciências - Bioquímica)- Universidade Federal do Paraná. Departamento de Bioquímica, 2010.

WESTAWAY, E.G. et al. Ultrastructure of Kunjin virus-infected cells: colocalization of NS1 and NS3 with double-stranded RNA, and of NS2B with NS3, in virus-induced membrane structures. **J. Virol.**, v. 71, p. 6650-6661, 1997.

WORLD HEALTH ORGANIZATION. **Dengue: guidelines for diagnosis, treatment, prevention, and control**. New Edition.1 ed., Geneva, 2009.

_____. **Yellow fever: Fact sheet n° 100**. Mar 2014. Disponível em <http://www.searo.who.int/thailand/factsheets/fs0010/en/>. Acesso em: 01 set.2016.

WORLD HEALTH ORGANIZATION. **Dengue and severe dengue:** Fact sheet n° 117. May 2015. Retrieved 3 February 2016a. Disponivel em <<http://www.who.int/mediacentre/factsheets/fs117/en/>>. Acesso em: 28 ago. 2016.

_____. **Chikungunya:** Fact sheet n° 327. July 2016b. Disponivel em <http://www.who.int/mediacentre/factsheets/fs327/en/>. Acesso em 03 set. 2016.

WINKLER, G. et al. Newly synthesized dengue-2 virus nonstructural protein NS1 is a soluble protein but becomes partially hydrophobic and membrane-associated after dimerization. **Virology**, v. 171, p. 302-305, 1989.

WRIGHT, P.J. et al. Definition of the carboxy termini of the three glycoproteins specified by dengue virus type 2. **Virology**. v. 171, n. 1, p. 61-67, 1989.

ANEXO

Tabela 1 – Genótipos selecionados para as sequências referências dos DENV, ZIKV, CHIKV e YFV

Genótipo do ZIKV	Número de Acesso	País	Ano
Africano	AY632535	Uganda	1947
Africano	KF383116	Senegal	1968
Africano	HQ234500	Nigéria	1968
Africano	KF383115	República Centro-Africana	1968
Africano	HQ234501	Senegal	1984
Africano	KF383117	Senegal	1997
Africano	KF383118	Senegal	2001
Africano	KF383119	Senegal	2001
Africano	KF383121	Senegal	-N/A-
Africano	KF268950	República Centro-Africana	-N/A-
Africano	KF268949	República Centro-Africana	-N/A-
Asiático	HQ234499	Malásia	1966
Asiático	EU545988	Micronésia	2007
Asiático	JN860885	Camboja	2010
Asiático	KF993678	Canadá	2013
Asiático	KJ776791	Polinésia Francesa	2013
Asiático	KU707826	Brasil	2015
Divergente do Oeste Africano	KF383120	Senegal	-N/A-
Outgroup	Spondweni	Spondweni virus	-N/A-
Genótipo do CHIKV	Número de Acesso	País	Ano
Asiático	HM045813	Índia	1963
Asiático	EF027140	Índia	1963
Asiático	EF027141	Índia	1973
Asiático	HM045790	Filipinas	1985
Asiático	FN295483	Malásia	2006
Asiático	FJ807897	Taiwan	2007
ESCA_IOC	HM045811	Tanzânia	1953
ESCA_IOC	HM045821	Senegal	1963
ESCA_IOC	AM258993	Reunião	2005
ESCA_IOC	AM258991	Seicheles	2005
ECSA_IN	AB455494	Japão	2006
ESCA_IOC	EF012359	Mauri	2006
ECSA_IN	EU244823	Itália	2007
ECSA_IN	FJ445426	Sri Lanka	2008
ECSA_IN	FN295485	Malásia	2008
ECSA_IN	GU199352	China	2008
ECSA_IN	GU301781	Tailândia	2009
ESCA_IOC	HM045784	República Centro-Africana	-N/A-
ESCA_IOC	HM045822	República Centro-Africana	-N/A-
ESCA_IOC	HM045792	África do Sul	-N/A-
Oeste Africano	HM045786	Nigéria	1964
Oeste Africano	HM045785	Senegal	1966
Oeste Africano	HM045815	Senegal	1979
Oeste Africano	HM045817	Senegal	2005
Oeste Africano	HM045818	Costa do Marfim	-N/A-
Oeste Africano	HM045820	Costa do Marfim	-N/A-

Genótipo do DENV-1	Número de Acesso	País	Ano
1I	AF350498	-N/A-	1980
1I	AY732478	Tailândia	1991
1I	AY732480	Tailândia	1994
1I	GQ868637	Camboja	2000
1I	AY732482	Tailândia	2001
1I	FJ469907	Singapura	2003
1I	GQ199835	Vietnã	2005
1I	FJ176779	China	2006
1I	AB608786	Taiwan	2008
1I	GU131895	Camboja	2009
1I	HQ891316	Sri Lanka	2009
1I	AY726552	Myanmar	2012
1I	AF298808	Djibuti	-N/A-
1II	AF180817	-N/A-	-N/A-
1III	AY713473	Myanmar	1971
1III	AY722801	Myanmar	1976
1III	AY722802	Myanmar	1996
1III	AY722803	Myanmar	1998
1IV	EF032590	-N/A-	1995
1IV	AB189121	Indonésia	1998
1IV	DQ672564	Estados Unidos	2001
1IV	DQ672561	Estados Unidos	2001
1IV	FJ196842	China	2003
1IV	GQ868602	Filipinas	2004
1IV	AB204803	Japão	2004
1IV	JN697056	Malásia	2005
1IV	U88535	-N/A-	-N/A-
1IV	AB074761	-N/A-	-N/A-
1IV	AB189120	Indonésia	-N/A-
1V	FJ205875	Estados Unidos	1995
1V	AF311956	-N/A-	1997
1V	EU482567	Estados Unidos	1998
1V	AB519681	Brasil	2001
1V	DQ285559	Reunião	2004
1V	HQ166037	México	2008
1V	KJ189351	Porto Rico	2012
1V	AY277666	-N/A-	-N/A-
1V	AY206457	-N/A-	-N/A-
1V	AY277665	-N/A-	-N/A-
Genótipo do DENV-2	Número de Acesso	País	Ano
2I (American)	EU056812	Porto Rico	1977
2I (American)	EU056811	Peru	1995
2I (American)	AF100469	-N/A-	-N/A-
2II (Cosmopolitan)	EU081180	Singapura	2005
2II (Cosmopolitan)	EU081179	Singapura	2005
2II (Cosmopolitan)	EU081177	Singapura	2005
2II (Cosmopolitan)	EU179859	Brunei	2006
2II (Cosmopolitan)	KC762660	Indonésia	2007

2II (Cosmopolitan)	KC762669	Indonésia	2007
2II (Cosmopolitan)	KC762680	Indonésia	2010
2II (Cosmopolitan)	KM279597	Singapura	2012
2III (SE Asian-America)	EU482582	Estados Unidos	1989
2III (SE Asian-America)	GQ868540	Venezuela	1990
2III (SE Asian-America)	GQ398290	Puerto Rico	1994
2III (SE Asian-America)	AY702036	Cuba	1997
2III (SE Asian-America)	AB122020	Republica Dominicana	2001
2III (SE Asian-America)	FJ898461	Belize	2002
2III (SE Asian-America)	EU687216	Estados Unidos	2005
2III (SE Asian-America)	EU482726	Estados Unidos	2006
2III (SE Asian-America)	GQ199868	Nicarágua	2007
2III (SE Asian-America)	HQ999999	Guatemala	2009
2III (SE Asian-America)	AF489932	-N/A-	-N/A-
2III (SE Asian-America)	M20558	-N/A-	-N/A-
2IV (Asian II)	KF744406	Filipinas	1995
2IV (Asian II)	KF744407	Filipinas	1996
2IV (Asian II)	HQ891023	Taiwan	2008
2IV (Asian II)	AF204177	China	-N/A-
2IV (Asian II)	AF038403	-N/A-	-N/A-
2V (Asian I)	DQ181806	Tailândia	1974
2V (Asian I)	DQ181805	Tailândia	1979
2V (Asian I)	DQ181804	Tailândia	1984
2V (Asian I)	DQ181802	Tailândia	1988
2V (Asian I)	KC964095	China	1998
2V (Asian I)	DQ181798	Tailândia	1999
2V (Asian I)	DQ181797	Tailândia	2001
2V (Asian I)	FM210211	Vietnã	2003
2V (Asian I)	GU131896	Camboja	2007
2VI (Sylvatic)	EF105387	Nigéria	1966
2VI (Sylvatic)	EF105379	Malásia	1970
2VI (Sylvatic)	EF105382	Burkina Faso	1980
2VI (Sylvatic)	EF105389	Senegal	1999
2VI (Sylvatic)	FJ467493	Malásia	2008
Genótipo do DENV-3	Número de Acesso	País	Ano
3I	AY858039	Indonésia	1998
3I	KC762682	Indonésia	2007
3I	KC762681	Indonésia	2007
3I	KC762686	Indonésia	2007
3I	KC762684	Indonésia	2007
3I	KC762687	Indonésia	2008
3I	KC762689	Indonésia	2008
3I	KC762688	Indonésia	2008
3I	KC762690	Indonésia	2008
3I	KC762685	Indonésia	2008
3I	KC762691	Indonésia	2008
3I	AY858037	Indonésia	-N/A-
3II	AY876494	Tailândia	1994
3II	AY923865	Tailândia	1994

3II	AY766104	Singapura	1995
3II	DQ675528	Taiwan	1998
3II	DQ675525	Taiwan	1998
3II	DQ675532	Taiwan	1998
3II	AY496871	Bangladesh	2002
3II	AY496877	Bangladesh	2002
3II	AY496874	Bangladesh	2002
3II	AY496873	Bangladesh	2002
3III	JQ411814	Sri Lanka	1989
3III	DQ675533	Taiwan	1999
3III	AY099337	Martinica	1999
3III	FJ639747	Venezuela	2000
3III	FJ547071	Estados Unidos	2000
3III	FJ898458	Peru	2002
3III	AY679147	Brasil	2002
3III	EU529702	Estados Unidos	2003
3III	FJ898442	México	2007
3III	GQ868578	Colômbia	2007
3III	AY099336	Sri Lanka	-N/A-
3III	AY770511	Índia	-N/A-
3V	EF629370	Brasil	2002
3V	AF317645	China	-N/A-
Genótipo do DENV-4	Número de Acesso	País	Ano
4I	GQ868594	Filipinas	1956
4I	AY618991	Tailândia	1977
4I	FJ196850	China	1990
4I	AY618992	Tailândia	2001
4I	KF041260	Paquistão	2009
4I	JQ513345	Brasil	2011
4II	GU289913	Colômbia	1982
4II	JF262782	Haiti	1994
4II	AY762085	-N/A-	1995
4II	JF262781	Venezuela	1995
4II	GQ252675	Estados Unidos	1995
4II	FJ024476	Colômbia	1997
4II	FJ882581	Venezuela	2007
4II	JN983813	Brasil	2010
4II	AF326573	-N/A-	-N/A-
4III	AY618988	Tailândia	1997
4III	AY618989	Tailândia	1997
4IV	JF262780	Malásia	1973
4IV	EF457906	Malásia	1975
Genótipo do Vírus Febre Amarela	Número de Acesso	País	Ano
South America I	TVP17398	Equador	1979
South America I	TVP14397	Equador	1977
South America I	TVP17402	Panamá	1974
South America I	TVP17435	Venezuela	1961
South America I	TVP17436	Venezuela	1959
South America I	JF912180	Brasil	1981

South America I	TVP17433	Trinidad e Tobago	1979
South America I	TVP17428	Trinidad e Tobago	1981
South America I	JF912187	Brasil	2000
South America I	JF912188	Brasil	2000
South America I	JF912189	Brasil	2001
South America I	TVP17438	Venezuela	1998
South America I	JF912190	Brasil	2002
South America I	TVP17431	Trinidad e Tobago	2009
South America I	TVP17426	Trinidad e Tobago	2009
South America I	TVP17434	Trinidad e Tobago	2009
South America I	TVP17429	Trinidad e Tobago	1954
South America I	TVP17394	Colômbia	1979
South America I	JF912182	Brasil	1984
South America I	JF912185	Brasil	1992
South America I	JF912179	Brasil	1980
South America I	JF912184	Brasil	1987
South America I	JF912186	Brasil	1994
South America I	JF912183	Brasil	1984
South America I	TVP17393	Brasil	1935
South America II	JF912181	Brasil	1983
South America II	TVP17409	Peru	1995
South America II	TVP17405	Peru	1995
South America II	TVP17425	Peru	1998
South America II	TVP17423	Peru	1998
South America II	TVP17420	Peru	1981
South America II	TVP17404	Peru	1996
South America II	TVP17389	Bolívia	2006
South America II	TVP17416	Peru	2007
South America II	TVP17388	Bolívia	1999
South America II	TVP17390	Bolívia	1999
South America II	TVP17387	Bolívia	1999
South America II	TVP17391	Bolívia	1999
South America II	TVP17421	Peru	1999
South America II	TVP17424	Peru	1999
South America II	TVP17410	Peru	1995
South America II	TVP17411	Peru	1995
South America II	TVP17419	Peru	1977
South America II	TVP17400	Equador	1997
South America II	TVP17392	Bolívia	1999
West African	AF094612	Trinidad e Tobago	1979
West African	JX898871	Senegal	1995
West African	JX898871	Senegal	1995
West African	AY640589	Gana	1927
West African	JX898875	Senegal	2000
West African	JX898874	Senegal	2000
West African	JX898873	Senegal	2000
West African	AY572535	Gambia	2001
West African	AY603338	Costa do Marfim	1999
West African	JX898868	Senegal	1995

West African	JX898870	Senegal	1996
West African	JX898876	Senegal	2001
West African	JX898878	Senegal	2005
West African	JX898880	Senegal	2005
West African	JX898877	Senegal	2005
West African	JX898869	Costa do Marfim	1973
West African	YFU54798	Costa do Marfim	1982
East African	AY968064	Angola	1971
East African	AY968065	Uganda	1948
East African	DQ235229	Etiópia	1961
East African	JN620362	Uganda	2010