

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas e Departamento de Biologia Geral
Programa Interunidades de Pós-Graduação em Bioinformática

Vagner de Souza Fonseca

**DEVELOPMENT OF BIOINFORMATICS TOOLS FOR ASSEMBLY AND
GENOMIC CHARACTERIZATION OF EMERGING AND RE-EMERGING
VIRUSES CIRCULATING IN BRAZIL**

Belo Horizonte
2022
Vagner de Souza Fonseca

**DEVELOPMENT OF BIOINFORMATICS TOOLS FOR ASSEMBLY AND
GENOMIC CHARACTERIZATION OF EMERGING AND RE-EMERGING
VIRUSES CIRCULATING IN BRAZIL.**

Versão Final

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Bioinformática.

Orientador: Dr. Luiz Carlos Júnior Alcantara

Coorientador: Dr. Tulio de Oliveira

Belo Horizonte

2022

043

Fonseca, Vagner de Souza.

Development of bioinformatics tools for assembly and genomic characterization of emerging and re-emerging viruses circulating in Brazil [manuscrito] / Vagner de Souza Fonseca. – 2022.

136 f. : il. ; 29,5 cm.

Orientador: Dr. Luiz Carlos Júnior Alcantara. Coorientador: Dr. Tulio de Oliveira.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Genômica. 3. Filogenia. 4. Arbovirus. I. Alcantara, Luiz Carlos Júnior. II. Oliveira, Tulio de. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004



UNIVERSIDADE FEDERAL DE MINAS GERAIS
 INSTITUTO DE CIÊNCIAS BIOLÓGICAS
 PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DA DEFESA DE TESE

VAGNER DE SOUZA FONSECA

Às oito horas e trinta minutos do dia **02 de maio de 2022**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Development Of Bioinformatics Tools For Assembly And Genomic Characterization Of Emerging And Re-emerging Viruses Circulating In Brazil**", requisito para obtenção do grau de Doutor em **Bioinformática**. Abrindo a sessão, o Presidente da Comissão, **Dr. Luiz Carlos Júnior Alcântara**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	Instituição	Indicação
Dr. Luiz Carlos Júnior Alcântara	FIOCRUZ Bahia	Aprovado
Dr. Aristóteles Góes Neto	UFMG	Aprovado
Dra. Raquel Cardoso de Melo Minardi	UFMG	Aprovado
Dr. Antonio Ricardo Khouri Cunha	FIOCRUZ Bahia	Aprovado
Dra. Ana Maria Bispo de Filippis	FIOCRUZ RJ	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 02 de maio de 2022.



Documento assinado eletronicamente por **Ana Maria Bispo de Filippis, Usuário Externo**, em 02/05/2022, às 10:38, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Antonio Ricardo Khouri Cunha, Usuário Externo**, em 02/05/2022, às 11:21, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Aristoteles Goes Neto, Coordenador(a) de curso de pós-graduação**, em 03/05/2022, às 09:33, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Raquel Cardoso de Melo Minardi, Professora do Magistério Superior**, em 03/05/2022, às 09:38, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Luiz Carlos Junior Alcantara, Usuário Externo**, em 11/05/2022, às 12:18, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1416514** e o código CRC **3D83F87B**.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisors and mentors, Dr. Luiz Alcantara, Prof. Tulio de Oliveira, and Dr. Marta Giovanetti, for their mentorship throughout my research and PhD training. Their indispensable contribution to my career is greatly appreciated.

I also extend my thanks to the staff and registrars at the Department of Bioinformatics, of Universidade Federal de Minas Gerais, for their support throughout my research work. Thanks to all the laboratory and data analysts and staff at the Flavivirus Laboratory of Oswaldo Cruz Foundation (LABFLA/IOC/Fiocruz) and Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP) for their support in the implementation of this work, and for the fantastic facilities, training and collaborations that I received throughout my studies. I would also like to acknowledge financial support from the Coordination of Superior Level Staff Improvement (CAPES).

Many thanks to my family, especially my mother, Mrs. Neide de Souza Fonseca, my parents-in-law, Mr. Antonio Teixeira and Teresa Elisabete Teixeira, for their unwavering support throughout my studies, from the first day of school up to now, my siblings, Aecio Brito Teixeira, Saulo Brito Teixeira, Verusca Bastos, Tiago Fonseca, and Mariana Souza, and most importantly my wife, Mrs. Simara Teixeira, for all her encouragement and support in every aspect of my life. Also, I thank my great friends on the scientific journey, Emmanuel James San, Joilson Xavier, and Talita Adelino.

Praise to God for blessing me with life, knowledge, guidance and health throughout the years of my studies. To God be all the glory.

RESUMO

As emergências e reemergências recentes de arbovírus como os vírus Chikungunya (CHIKV), Zika (ZIKV), Dengue (DENV), Febre Amarela (YFV), Mayaro (MAYV), Oropouche (OROV) e Febre do Nilo Ocidental (WNV) no Brasil ilustram a necessidade de monitoramento genômico rápido para que contramedidas possam ser prontamente organizadas.

O objetivo geral deste estudo foi, portanto, melhorar a qualidade dos serviços de saúde pública, por meio de um monitoramento ativo de vírus circulantes e co-circulantes por análise genômica e de bioinformática que são necessárias para: (i) identificar possíveis novos vírus emergentes, bem como identificar aqueles já circulantes mas com baixa viremia; ii) reduzir a subnotificação de casos de coinfeção; iii) compreender a dinâmica de disseminação espaço-temporal de possíveis vírus circulantes e co-circulantes, e (iv) determinar os fatores que afetam a disseminação e evolução clínica das infecções causadas por arbovírus.

Ferramentas de bioinformática eficientes capazes de analisar o grande volume de dados gerados a partir do sequenciamento de alto rendimento são essenciais para entender a epidemiologia das infecções causadas por patógenos virais em uma determinada área. Essas ferramentas também podem ser usadas para fornecer dados oportunos sobre a disseminação de infecções para outras regiões geográficas. Nesse contexto, estratégias criativas de bioinformática podem democratizar e descentralizar o processo de gerenciamento e análise de dados, permitindo a realização de um sistema de vigilância global automatizado, em tempo real e de acesso aberto.

Os dados genômicos gerados neste projeto permitiram aumentar a compreensão sobre a disseminação geográfica e temporal dos vírus circulantes. Isso pode ter influenciado políticas públicas melhorando as medidas para controlar epidemias, monitorar a dinâmica e disseminação de novas variantes virais com consequente implementação de programas de controle mais eficientes para vigilância genômica em tempo real da síndrome febril aguda relacionada a arbovírus.

Palavras-Chave: Arbovirus, Bioinformática, MinION, Filogenia e NGS.

ABSTRACT

The recent emergence and re-emergence of arboviruses such as Chikungunya (CHIKV), Zika (ZIKV), Dengue (DENV), Yellow Fever (YFV), Mayaro (MAYV), Oropouche (OROV), and West Nile Fever (WNV) viruses in Brazil illustrate the need for rapid genomic monitoring so that countermeasures can be readily organized.

The overarching aim of this study was therefore, to improve the quality of public health care services, through an active monitoring of circulating and co-circulating viruses by genomics and bioinformatics analysis. These are necessary: (i) to identify possible new emerging viruses, as well as to identify those already circulating but low in viraemia; ii) to reduce underreporting of co-infection cases; iii) to understand the dynamics of spatiotemporal dissemination of possible circulating and co-circulating viruses, and (iv) to determine the factors affecting the spread and clinical outcome of infections caused by arboviruses.

Efficient bioinformatics tools capable of analysing the large volume of data generated from high throughput sequencing are essential to understand the epidemiology of infections caused by viral pathogens in a given area. These tools may also be used to provide timely data about the spread of infections to other geographical regions. In this context, creative bioinformatics strategies can democratize and decentralize the data management and analysis process, allowing the realization of an automated, real time and open access global surveillance system.

Indeed, the genomic data generated in this project improved understanding of geographic and temporal spread of circulating viruses. This likely influenced policy and improved measures to control epidemics, monitor the dynamics and spread of new viral strains, and led to the implementation of more efficient control programs for real-time genomic surveillance of acute febrile syndrome related to arboviruses.

Keywords: Arbovirus, Bioinformatics, MinION, Phylogeny and NGS.

LIST OF FIGURES

Figure 1. Phylogenetic reconstruction of the Flavivirus genus. (Reproduced with modifications from Cook et al., 2012)	15
Figure 2. Flavivirus genome. (Reproduced from Viral Zone: Flavivirus; https://viralzone.expasy.org/24?outline=all_by_species).....	16
Figure 3. Chikungunya virus genome. (Reproduced from ViralZone: Togaviridae; https://viralzone.expasy.org/3)	18

ABBREVIATIONS

AGA	Annotated Genome Aligner
BAM	Business Activity Monitoring
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BWA	Burrows-Wheeler Aligner
CHIKV	Chikungunya Virus
CONEP	Comissão Nacional de Ética em Pesquisa, Ministério da Saúde
DENV	Dengue Virus
DNA	Deoxyribonucleic Acid
ECSA	East-Central-South-African
ISF	Insect-Specific Flaviviruses
JEV	Japanese Encephalitis Virus
LACENS	Central Public Health Laboratories (Laboratórios Centrais de Saúde Pública)
LRT	Likelihood Ratio Test
MAFFT	Multiple Alignment using Fast Fourier Transform
MAYV	Mayaro Virus
MBV	Mosquito-Borne Viruses
ML	Maximum Likelihood
mRNA	Messenger Ribonucleic Acid
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NIH	National Institutes of Health
NJ	Neighbor Joining
No Known Vector Viruses	No Known Vector Viruses
NTR	Non-Translatable Region
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
OROV	Oropoche Virus
PAML	Phylogenetic Analysis by Maximum Likelihood
PAUP*	Phylogenetic Analysis using Parsimony (*and other methods)
PCR	Polymerase Chain Reaction
PhyML	Fast and accurate estimation of Phylogenies using Maximum Likelihood
QIAGEN	QIAmp Viral RNA Minikit
RDP	Detection and Analysis of Recombination
RdRp	RNA-dependent RNA Polymerase
RNA	Ribonucleic Acid
RT-qPCR	Real-Time Quantitative Polymerase Chain Reaction
SIFT	Sorting intolerant from tolerant
SNP	Single Nucleotide Polymorphism
SQK-LSK109	Ligation Sequencing Kit
TAR-VIR	Targeted Viral
TBV	Tick-Borne Viruses
UTR	Untranslated Region
VIP	Virus identification Pipeline
ViPR	Virus Pathogen Database and Analysis Resource

WNV
YFV
ZIKV

West Nile Virus
Yellow Fever Virus
Zika Virus

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW.....	13
1.1 Background	13
1.2 Literature Review.....	14
1.2.1 Arboviruses.....	14
1.2.1.1 Flavivirus.....	14
1.2.1.1.1 Dengue Virus	16
1.2.1.1.2 Zika Virus	17
1.2.1.2 Alphavirus	17
1.2.1.2.1 Chikungunya Virus	18
1.2.2 Coronavirus.....	19
1.2.2.1 Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2)	20
1.2.3 Genomic Surveillance and Bioinformatics Tools.....	20
1.3 Justification	42
1.4 Main objective.....	42
1.4.1 Specific objectives	43
1.5 Ethical approval	43
CHAPTER 2: GENOME DETECTIVE: AN AUTOMATED SYSTEM FOR VIRUS IDENTIFICATION FROM HIGH-THROUGHPUT SEQUENCING DATA.....	44
CHAPTER 3: A COMPUTATIONAL METHOD FOR THE IDENTIFICATION OF DENGUE, ZIKA AND CHIKUNGUNYA VIRUS SPECIES AND GENOTYPES	48
CHAPTER 4: GENOME DETECTIVE CORONAVIRUS TYPING TOOL FOR RAPID IDENTIFICATION AND CHARACTERIZATION OF NOVEL CORONAVIRUS GENOMES	64
CHAPTER 5: WEST NILE VIRUS IN BRAZIL	69
CHAPTER 6: SYNTHESIS OF RESEARCH FINDINGS.....	84
7.1 Key Themes	84
7.2 Recommendations for policy	86
7.3 Publications Declaration	87
7.3.1 Papers.....	87
7.3.2 Book.....	89
7.3.3 Chapter.....	90
7.4 References	91
APPENDICES.....	98
Appendix 1 Supplementary material to the manuscript entitled “Genome Detective: an automated system for virus identification from high-throughput sequencing data”	98
Appendix 2 Supplementary material to the manuscript entitled “A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes”	109
Appendix 3 Supplementary material to the manuscript entitled “Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes”	117
Appendix 4 Supplementary material to the manuscript entitled “West Nile Virus in Brazil”	131

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Background

Arboviruses are diseases transmitted by the bite of arthropod vectors, mostly mosquitoes and ticks. Historically, arbovirus research has been of low global priority, mainly because it is mostly confined to under-developed countries in the tropics. However, the 21st century has witnessed a rapid and worrying re-emergence of these viruses, which is linked to factors such as virus adaptation to new vectors, world population growth, urbanization of forest areas, climate change, globalization, and ease of population displacement [1]. Among the main arboviruses of medical importance are the Dengue virus (DENV), Chikungunya virus (CHIKV), and the Zika virus (ZIKV), whose main vector is mosquitoes of the genus *Aedes*, such as *Aedes aegypti* and *Aedes albopictus*. In Brazil, these three viruses currently circulate simultaneously. They are responsible for an extremely worrying epidemiological crisis, placing the country at the centre of global attention with regard to the threat of arboviruses.

Diseases caused by arboviruses are associated with significant epidemics and, as a consequence, diagnosis and treatment place a significant financial burden on the public health system [2]. Clinical diagnosis is considered complex due to the similarity of symptoms to other diseases, serological cross-reactions, the presence of clinically asymptomatic or oligosymptomatic disease, and the difficulty of accessing reference laboratories that can perform a molecular and/or differential serological diagnosis [1,3].

DENV is already endemic in Brazil and has caused numerous epidemics. In 2014-2015 there was the introduction of CHIKV and ZIKV. In addition, the emergence of other arboviruses which were previously circulating in the Amazon and in tropical areas of Central America and the Caribbean, such as the Mayaro virus (MAYV) and Oropoche virus (OROV), have recently been reported in Northeast, Southeast, and Midwest Brazil. This raises serious concerns about public health in the country [4].

The epidemiological cycle of MAYV and OROV is similar to that of wild-type YFV and takes place with the participation of wild mosquitoes, mainly of the genus *Haemagogus*. Other genera of mosquitoes, such as *Culex*, *Sabethes*, *Psorophora*, *Coquillettidia*, and *Aedes*, may participate in the maintenance cycle of these viruses as they can amplify and maintain these viruses in their natural environment [5–7].

The sequencing of viral genomes has played an important role in the fight against emerging, and re-emerging epidemics since the use of bioinformatics tools and the combination of viral and epidemiological genomic data has generated useful information for understanding the past and future of circulating arboviruses [8,9].

In terms of real-time active genomic surveillance, recent successes in sequencing Ebola, Zika, and YFV using the nanopore sequencing technology present in the MinION handheld device [10–12] have proven that high quality complete viral genome sequences can be generated in near real-time (often < 2 business days) during viral outbreak events. The portability and real-time data production provided by MinION allows genomic data to be generated at the source of origin of the outbreak in real-time, allowing fast data channelling, and consequently, quick intervention strategies.

As can be seen from the above background, it is necessary to track transmission patterns and geographical spread over time to determine the emergence and re-emergence of circulating and co-circulating arboviruses.

1.2 Literature Review

1.2.1 Arboviruses

Arboviruses comprise a large group of zoonotic viruses that infect hematophagous arthropods and are commonly transmitted to humans through mosquito bites. Arboviruses are classified into four main families: *Togaviridae* (genus *Alphavirus*), *Flaviviridae* (genus *Flavivirus*), *Bunyaviridae* (genus *Orthobunyavirus* and *Phlebovirus*), and *Reoviridae*. Most arboviruses have a single-stranded Ribonucleic Acid (RNA) genome with spherical morphology and a diameter that varies between 45-120 nanometers [13,14].

It is estimated that there are more than 545 species of arboviruses, more than 150 of which are related to diseases in humans. Most of which are caused by new agents or known agents that affect sites and species that have not yet experienced the disease. They are kept in a transmission cycle between arthropods (vectors) and vertebrate reservoirs as the main amplifier hosts [15,16]. The *Flavivirus* and *Alphavirus* genera account for more than 90% of infections caused by arboviruses in tropical countries.

1.2.1.1 Flavivirus

The Flavivirus genus comprises 53 different species of virus, harbouring more than 70 described viruses [17–20]. The word ‘flavivirus’ is derived from the Latin word ‘*flavus*’, which means yellow, due to jaundice (a condition that causes a yellowish skin colour) caused by the YFV, the prototype of the family [21]. The vast majority of these are pathogens transmitted by arthropods: mosquitoes transmit 27 species of viruses, ticks transmit 12, whilst 14 species do not have their vectors identified yet [22]. Symptoms of infection range from mild fever and malaise to fatal encephalitis (acute brain infections caused by a virus, bacteria, fungi or parasites and even chemical or toxic substances), and haemorrhagic fever [23].

The classification of viral species is based genomic organization, vector association, morphology, viral ecology, and the relationship of nucleotide sequences. Flaviviruses can cause epidemics with high mortality and morbidity rates, hence the importance of research. The Flaviviruses that most infect human hosts are DENV, West Nile Virus (WNV), Japanese Encephalitis Virus (JEV), and ZIKV [24].

The construction of phylogenetic inference analyses based on alignments of multiple nucleotide (nt) sequences of viruses belonging to this genus indicate the existence of four large monophyletic groups (Figure 1). The coding sequences used in the analysis of this work [25] were collected from the RNA Virus Database [26]. The first group of sequences, represented by the colour blue, has its vector transmitted by the mosquito, known in literature as Mosquito-Borne Viruses (MBV). The second group of sequences, represented by the colour red, has its vector transmitted by ticks, known in literature as Tick-Borne Viruses (TBV). The third group of sequences, represented by the colour orange are not yet associated with a known an arthropod and are therefore referred to as No Known Vector Viruses (NKVV). Finally, the fourth group of sequences, represented by the colour green, are sequences transmitted by specific insects. These are termed Insect-Specific Flaviviruses (ISF) in literature [25].

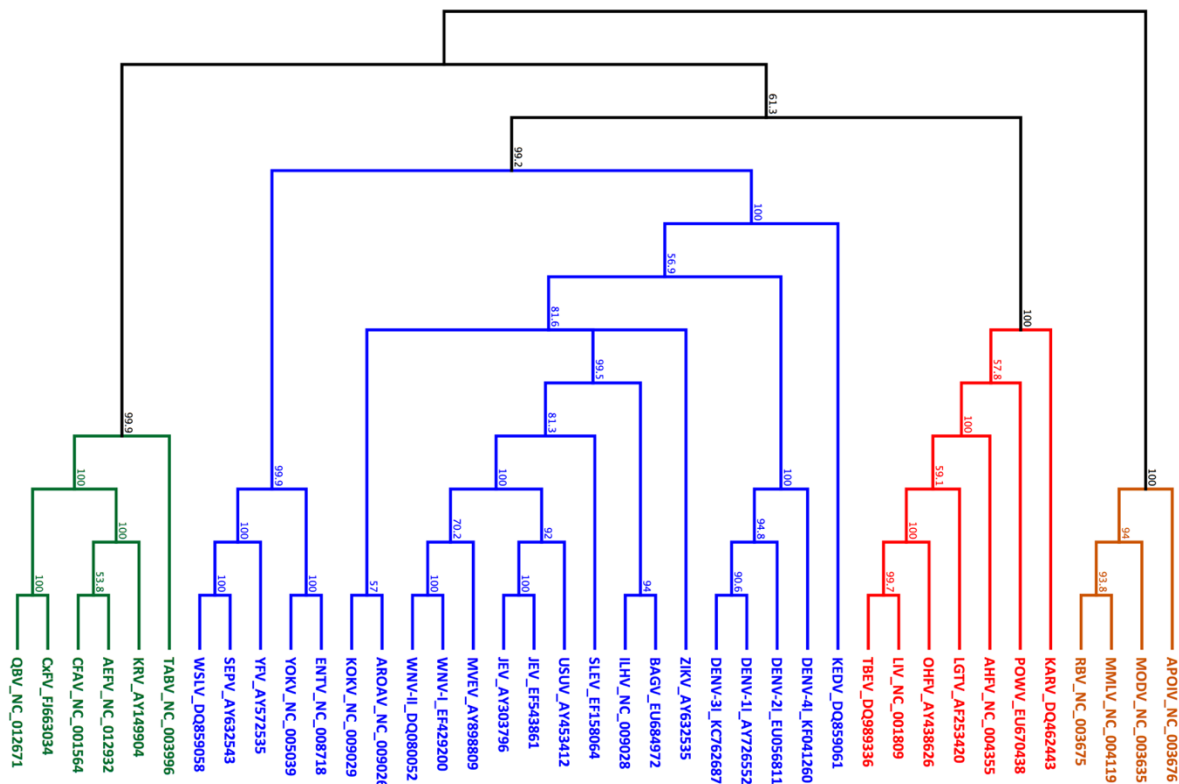


Figure 1. Phylogenetic reconstruction of the Flavivirus genus. (Reproduced with modifications from Cook et al., 2012)

The monophyletic group formed by the ISF constitutes an ancestral lineage of the genus *Flavivirus* in phylogenetic trees [23]. ISFs were initially designated as mosquito-specific viruses, but after identifying the virus in sandflies (Diptera and Psychodidae insects), this last nomenclature was invalidated and called ISF [27].

The genome of the *Flavivirus* genus is composed of a single strand of RNA with approximately 11kb and positive polarity, which flanks a single Open Reading Frame (ORF) encoding a polyprotein with about 3,400 amino acid residues. After its synthesis, this polyprotein is processed by viral and cellular proteases with three structural proteins: capsid [C], membrane [M] and envelope [E]; and seven non-structural proteins (NS), called NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5 [22,28–30]. In addition, at both ends of the genome, there are two Untranslated Region (UTR) called UTR-5' and UTR-3', which tend to form secondary structures that perform regulatory functions and virus expression, such as replication, virulence, and pathogenicity [31] (Figure 2).

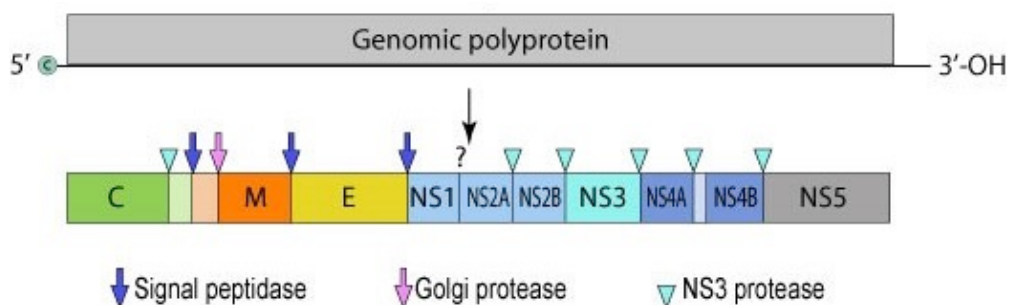


Figure 2. Flavivirus genome. (Reproduced from Viral Zone: *Flavivirus*; https://viralzone.expasy.org/24?outline=all_by_species).

1.2.1.1.1 Dengue Virus

The first vectors isolated in Japanese soldiers were called DENV-1, the second, isolated in American soldiers, were called DENV-2. DENV-3 and DENV-4 were isolated in the Philippines [32]. Chinese manuscripts published during the Chin dynasty (AD 265 to 420) report a Dengue-like disease called "water poison". There are also reports from 1889 to 1990 of an epidemic similar to Dengue in Jakarta, Cairo, and Philadelphia [33]. Between the 1950s and 1960s, an outbreak of haemorrhagic fever occurred in Manila and Bangkok [31]. Before World War II, pandemics caused by Dengue occurred every 20 years, but they were not frequent in the same region [34]. Favourable conditions resulting from ecological changes and economic activities such as urbanization in Southeast Asia provided the proliferation of the mosquito vector, thus initiating a global DENV pandemic [35,36].

While the *Aedes aegypti* vector was eradicated from the Americas 50 to 70 years ago to control YFV in urban regions, the Dengue vector remained in Southeast Asia sustaining a large circulation of several serotypes of the Dengue haemorrhagic virus in the area [35,37]. From the 1980s onwards, the number of countries with a Dengue epidemic increased significantly as new genotypes emerged

[35]. *Aedes aegypti* mosquito was reintroduced in the American continent with new serotypes in susceptible populations, thus increasing vector transmission [38,39]. Dengue is now endemic in more than 100 countries, distributed across tropical Asia, Africa, Australia, Central America, and South America where it causes high rates of infection [37].

DENV is classified into four closely related serotypes, called: DENV-1, DENV-2, DENV-3 and DENV-4 [31] and 18 genotypes (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-II, 3-III, 3-IV, 4-I, 4-II, 4-III and 4-IV). This nomenclature consists is based on the annotations of the DENV genomes in Virus Pathogen Database and Analysis Resource (ViPR) [40].

1.2.1.1.2 Zika Virus

ZIKV was discovered in Uganda in the Zika Forest by researchers in 1947 studying YFV cycle [41]. The first ZIKV infection in humans occurred in Nigeria in 1954 [42]. Symptoms of this first infections were fever, headache, diffuse joint pain, and in one case, mild jaundice. The first infection detected by the *Aedes aegypti* vector occurred in Malaysia in 1966 [43]. After 11 years of suspicions of ZIKV infections on the Asian continent, Indonesia reported that seven patients had a fever, malaise, stomach pain, anorexia, and dizziness [44].

The first outbreaks of ZIKV infection occurred on an island in the Federated States of Micronesia in 2007. Fever, rash, conjunctivitis, and arthralgia were identified in 59 patients. Of these, 49 cases were positive for ZIKV [45]. In 2013, ZIKV reached French Polynesia and several islands in Oceania, where *Aedes aegypti* and *Aedes albopictus* are mostly found [46]. The outbreak in Polynesia infected approximately 10,000 people with fever, maculopapular rash, arthralgia, and conjunctivitis [47]. In 2014, new cases were also registered in New Caledonia and the Cook Islands.

To date, no deaths have been attributed to the ZIKV. However, in February 2014, Chilean public health authorities confirmed a case of autochthonous transmission of ZIKV infection on Easter Island [48]. In the first half of 2015, ZIKV was identified for the first time in the Americas, in some states in the Northeast region of Brazil. Since then, the virus has spread throughout the country and other countries in the Americas, with the exception of Chile and Canada [48].

ZIKV has two genotypes classified by phylogenetic analysis, these being African and Asian. The African genotype was identified in Ugandan patients in 1947 and many other African countries since then. The second group was first identified in Malaysia in 1966 and has caused epidemics in Micronesia, French Polynesia, and New Caledonia [43].

1.2.1.2 Alphavirus

The primary virus of the Alphavirus genus is CHIKV. The genome consists of a linear, single-stranded, positive polarity RNA molecule with approximately 11.8 kb. This viral genomic RNA resembles cellular messenger Ribonucleic Acid (mRNA) in that it has a cap structure at the 5' end and a polyA tail at the 3' end. The CHIKV genome has two ORFs (Figure 3). One of them occupies two-thirds of the 5' portion of the genome and encodes a polyprotein that, after proteolysis, gives rise to multifunctional non-structural proteins (nsP1, nsP2, nsP3 and nsP4), which form the viral *replicase*. The other ORF, separated from the first by a junction region, encodes a second polyprotein generated by proteolytic processing, the structural proteins [C, E1, PE2 (E3+E2), and 6K]. Non-Translatable Region (NTR) is flanked at the 5' and 3' ends by untranslated sequences, called 5'NTR and 3'NTR, respectively [49].

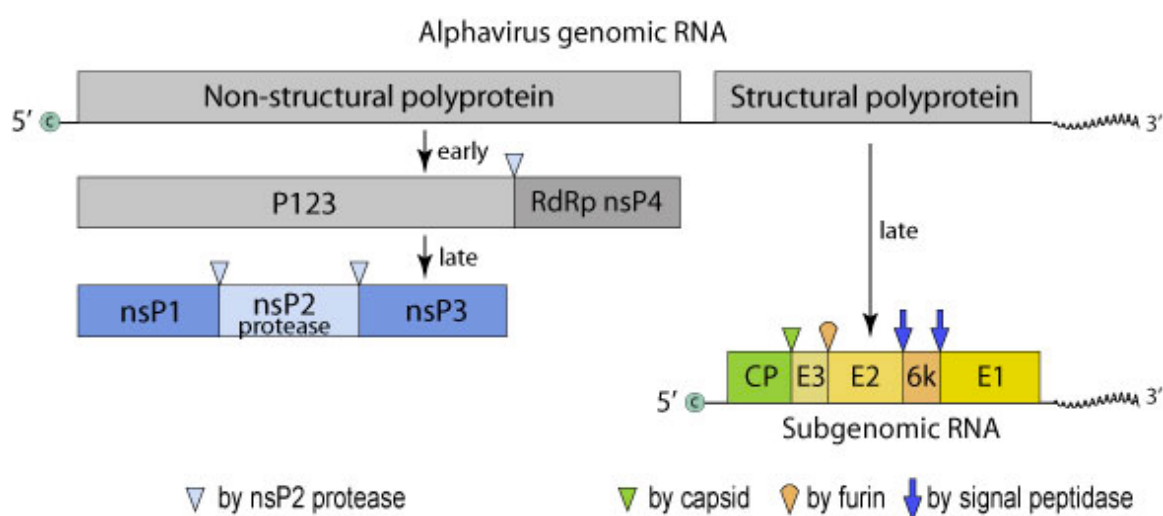


Figure 3. Chikungunya virus genome. (Reproduced from ViralZone: Togaviridae; <https://viralzone.expasy.org/3>)

1.2.1.2.1 Chikungunya Virus

CHIKV was isolated for the first time in blood samples obtained during an epidemic suggestive of Dengue-like Dengue, which occurred between 1952-1953 in Tanzania, a country located in East Africa [50,51]. The difficulty in walking caused by the intensity of the involvement of the joints served as inspiration for the name given to that disease: chikungunya, which in the Makonde dialect, spoken in the region, meant something like “walking bent over the body” [50,52]. Over the next 50 years, following its isolation, the circulation of CHIKV was restricted to Africa and Asia [53].

In 2004, CHIKV re-emerged during an epidemic in Kenya, reaching several islands in the Indian Ocean and India in the following years. During this period, hundreds of imported cases were identified in Europe, the Caribbean and North America [53]. In 2005, a large outbreak of CHIKV occurred on the islands of the Indian Ocean. Several Asian countries were affected in 2006 and 2007,

and more than 1.9 million people were infected [54]. Also, in 2007, Italy registered an outbreak with 197 reported cases transmitted by the vector of *Aedes albopictus* [54].

In October 2013, the first autochthonous cases were diagnosed on the island of Saint Martin, located in the so-called French Caribbean [55,56]. Since then, CHIKV infection has been confirmed in more than 43 countries, making this the first documented outbreak of CHIKV in the Americas. In 2014, around 1,300,000 suspected cases of CHIKV were registered in Caribbean islands, the United States and Latin American countries, leading to the death of around 191 people [54].

Three distinct groups of CHIKV capable of causing infection, were identified by phylogenetic analysis. The first was classified as a Central-East-South-African (ECSA) genotype, originating in Africa. The second genotype was classified as Caribbean Asian and the third a West African genotype, which is more divergent and less widespread than the previous two [57].

1.2.2 Coronavirus

Coronaviruses (CoV) belong to the Coronavirinae subfamily, Coronaviridae family. They are viruses with a single strand of RNA and a helical nucleocapsid, a structure composed of the virus's nucleic acid and its protein envelope, the capsid [58]. Its name is due to spicules, prominent structures, present on the virus's surface, which gives it the appearance of a corona [59].

Coronaviruses are viruses mainly related to respiratory and gastrointestinal tract infections, being taxonomically classified into four genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus [60]. Seven different species of CoV causing infection in humans have been described: HCOV-SARS, HCOV-OC43, HCOV-NL63, HCOV-MERS, HCOV-229e, HCOV-HKU1 and the newly discovered SARS-CoV-2 [61]. The group can also be divided into endemic and epidemic viruses according to the record of their discovery. The first record of infection Coronavirus in humans was reported in the 60s, when HCOV-OC43 and HCOV-229e were described as causing "colds", until then without any severe complications [62,63]. The next record of endemic coronaviruses that infected humans occurred in 2004-2005 when species HCOV-NL63 and HCOV-HKU1 were discovered. Up until this point, Coronaviruses were classified as a family of endemic viruses and associated with short colds being equated with the Influenza virus [64,65].

The first case of Severe Acute Respiratory Syndrome SARS was identified in Foshan, China, in November 2002. A few months later, in July 2003, the virus had already spread to more than 30 countries, causing about 8,000 infections and approximately 800 deaths. Nine years later, a new species of Coronavirus was identified in a lung sample from a 60-year-old patient who died of respiratory failure in Saudi Arabia, this being the first Middle East Respiratory Syndrome (MERS) notification. Coronavirus. In the 2012 epidemic in Saudi Arabia, MERS caused more than 2500 cases with a total of 861 deaths, with a fatality rate of 35% [66–68].

CoVs are enveloped viruses with positive polarity (+ssRNA) single-stranded RNA genetic material, with a genome of 26 to 32 kilobases. The RNA of coronaviruses contains multiple ORFs (open read frames). The largest ORF located at the 5' end occupies about 2/3 of the genome is composed of two overlapping ORFs called ORF1a and ORF1b, responsible for encoding its polyprotein [69]. The other major ORF is located at the 3' end and encodes four structural proteins: Spike (S), Membrane (M) and Envelope (E) and Nucleocapsid (N).

1.2.2.1 Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2)

Thought to have emerged from an animal host in late 2019, SARS-CoV-2 has subsequently spread to nearly every country in the world. The declaration of the disease caused by the novel coronavirus (COVID-19 - Coronavirus Disease 19) as a Public Health Emergency of International Concern and subsequent classification as a pandemic, on March 11, 2020, by the World Health Organization (WHO), has raised serious concerns around the world because of the rapid spread of the SARS-CoV-2 virus and the ability of hospitals to meet the high demand for hospitalizations [70–72].

The resulting COVID-19 pandemic has caused almost 340 million cases and 5 million deaths, strained local and global economies, and laid bare the racial, social, and gender inequities that dictate access to health resources. Since the start of the Anthropocene epoch, human led, or “anthropogenic” activities have exerted environmental effects on a range of scales, straining ecosystem services. Although the exact nature of the SARS-CoV-2 emergence is still uncertain, anthropogenic forces likely played a key role.

On February 11 2020, WHO held a meeting in Geneva with more than 300 scientists and public health experts to assess current knowledge about the new virus and discuss a plan to accelerate research on questions that need to be answered that can contribute to shorten the pandemic and prevent future outbreaks [73]. In the context of the WHO Plan of Action to Prevent Epidemics R&D Blueprint, gaps in knowledge were discussed and research priorities were identified, including understanding the natural history of the virus, monitoring the adaptation of the virus and understanding the transmission dynamics of the virus [73,74].

As the large-scale spread caused by COVID-19 and the recommendations suggested by the WHO of social distancing/isolation as measures to control and spread the virus [75], Brazilian states and municipalities followed these recommendations and, through government decrees, stipulated the suspension of face-to-face classes in schools, colleges and universities [76]. Concern about the ability of health systems to meet growing demand, including the need for hospital beds and respirators, has led to the proposition of measures to contain the rapid escalation of the disease [77].

1.2.3 Genomic Surveillance and Bioinformatics Tools

Genomic surveillance combines genomic and epidemiological data with bioinformatics tools that generate essential information for understanding the past and future of circulating human viruses. A continuous and structured system of viral genomics, epidemiology and bioinformatics, integrated with surveillance data, can provide timely data to inform adequate responses to emerging and re-emerging viruses [74].

To further clarify this topic, we have published the book chapter, "Mosquito-Borne Viral Diseases: Control and Prevention in the Genomics Era" from the book "Vector-Borne Diseases - Recent Developments in Epidemiology and Control", published by the publisher IntechOpen.

The chapter highlights how genomic surveillance has responded to the vast increase in information caused by the increased sophistication of bioinformatic tools.

Book Chapter**Mosquito-borne viral diseases: control and prevention in the genomics era**

Vagner Fonseca^{1,2}, Joilson Xavier², San Emmanuel James¹, Tulio de Oliveira¹, Ana Maria Bispo de Filippis³, Luiz Carlos Junior Alcantara^{2,3}, Marta Giovanetti^{2,3}.

¹KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZuluNatal, Durban 4001, South Africa; ²Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; ³Laboratório de Flavivírus, Instituto Oswaldo Cruz Fiocruz, Rio de Janeiro, Brazil.

Abstract

Mosquito-borne viral diseases are infections transmitted by the bite of infected mosquitoes. The burden of these diseases is highest in tropical and subtropical areas and they disproportionately affect the poorest populations. Since 2014, major outbreaks of dengue, chikungunya, yellow fever and zika have afflicted populations and overwhelmed health systems in many countries. Distribution of mosquito-borne diseases is determined by complex demographic, environmental and social factors, causing diseases to emerge in countries where they were previously unknown. Coupling genomic diagnostics and epidemiology to innovative digital disease detection platforms raises the possibility of an open, global, digital pathogen surveillance system. Considering pathogen surveillance in mind, real-time sequencing, bioinformatics tools and the combination of genomic and



epidemiological data from viral infections can give essential information for understanding the past and the future of an epidemic, making possible to establish an effective surveillance framework on tracking the spread of infections to other geographic regions.

Keywords: Mosquito-borne viral diseases; Arboviral infections; Genomics Epidemiology; Next-Generation Sequencing; Genomic Surveillance; Viral pathogens.

1. Introduction

Mosquito-borne viral diseases have lately integrated worldwide headlines since the emergence of arbovirus outbreaks in big urban areas. According to the World Health Organization more than 17% of all infectious diseases registered worldwide is represented by vector-borne diseases, and they account for more than 700,000 deaths annually [1]. Due to this scenario of increasing cases number and expansion to new areas, the spread of infectious diseases was listed second in the top 10 risks in term of impact according to the Global Risks 2015 report [2].

Mosquitos of the genus *Aedes* have been responsible for the emergence and re-emergence of many arboviral diseases worldwide [3]. The species *Aedes aegypti* is the main vector species responsible for the major arbovirus epidemics recorded in recent years [4]. The species *A. aegypti* and *A. albopictus* are possibly suitable to survive and establish in 215 countries/territories, and their expanding range is underlined by the increasing number of countries reporting transmission of mosquito-borne viruses. Transmissions of arboviruses, such as zika, dengue, chikungunya, yellow fever, and Rift Valley fever, have been reported in 85, 111, 106, 43, and 39 countries, respectively [5]. Projections indicated that 3,83 billion people are living in areas prone to transmission of dengue and it is predicted that by 2050 large increases in dengue suitability will be seen in southern



Africa and in the Sahel in West Africa [22]. Bhatt et al. (2013) projected the global burden of dengue around the world whose estimate indicated that 96 million dengue infections occur per year worldwide and this number represents infections that manifest at any level of the disease severity [6]. the Americas, comprising North and South America, registered more than 2 million dengue cases in 2016, and more than 1,4 million cases in 2019 [7]. For chikungunya fever, the Americas registered more than 94,000 cases in 2018, and in that same region, zika fever accounted for more than 650,000 cases in 2016 [8, 9]. High number of cases of arboviral diseases was also registered in other regions in recent years, such as in the western pacific region where more than 375,000 suspected dengue cases were reported in 2016 [10]. In Africa, the government of Congo reported 6,149 suspected cases of chikungunya until April 2019, and more than 13,000 chikungunya cases were reported in Sudan until October 21018 [11, 12]. The increasing in frequency and distribution of arboviral diseases in recent years represents a worrying burden not only for the public health system, but also for the economic sector [13]. Some estimates of the economic costs of arboviral infections have been made and for the case of dengue infections, it has been estimated that the median cost of of all reported dengue hospital admissions registered in a municipality from Brazil was US\$ 259.9 per hospitalization [18]. Also, in Maldives, in the Indian Ocean, dengue fever represented a total cost of \$3 million in 2015 [15]. Another estimate indicated that West Nile fever hospitalized cases in US represented a total cumulative cost of \$778 million between 1999 and 2012 [16].

Dengue and chikungunya are two arboviral diseases present in the list of neglected tropical diseases from the World Health Organization. Neglected tropical diseases are a group of diseases that have received insufficient public attention, thrive in tropical and subtropical areas, and strongly affect populations living in poverty [12]. It is argued that arboviruses can be considered a group of neglected tropical diseases, since they can have a long-lasting impact in the health and economic life of affected populations



[17]. Some studies have argued that socioeconomic factors and land-use changes associated with the effects of climate change and global travel, and trade modulate the dynamics of expansion of emerging e re-emerging mosquito-borne diseases [18, 19, 20, 21]. Movement of people between neighbouring countries have been considered a good predictor for chikungunya spread in the Caribbean and Indian Ocean [22]. The expansion of the geographic distribution of arbovirus has significant negative impact on public health in many regions of the world. As measures to reduce such impacts, it has been argued about the relevance to public health of the implementation of a surveillance system that monitors virus diffusion and the appearance of new genetic variants [23]. In this sense, the use of genomic sequencing data and bioinformatics have been employed in the the study of virus evolution, aiming to elucidate phylogenetic relationships and patterns of virus spread during an epidemic [24].

2. Genomic Surveillance

Infectious diseases continue to be one of the leading causes of death worldwide [25] and pathogens such as viruses can evolve and spread rapidly, leading to the emergence of newly-mutated human pathogens, more virulent strains, as well as antibiotic and drug resistant organisms [26, 27]. In this context, genomic surveillance aims are to: (i) to perform global surveillance of pathogens using whole genome sequencing; (ii) to understand drug resistance, emergence and spread of viral pathogens. Several approaches have been developed and are widely used for the quick detection and identification of viral pathogens (i.e., diagnostics). Some of them are based on different serological and molecular strategies including, for example, assays based on real-time polymerase chain reaction [28]. Even though these kinds of approaches present high sensitivity and specificity for their purpose, they are more suitable for diagnostics only and cannot provide detailed genomic information [29].



Bearing these limitations in mind, the main point of developing new genomic surveillance tools is to answer the following inquiry: what sort of questions are important for genomic surveillance that cannot be addressed by conventional RT-qPCR or serology? (i) RT-qPCR assays do not allow genotype classification, neither does it help identify particular and/or characteristic transmission routes; (ii) RT-qPCR assays also do not allow to determine how fast a viral pathogen is being transmitted and in what direction it is spreading; (iii) Serological and molecular assays also cannot help identify epidemiologically linked individuals, neither predict future outbreaks; (iv) finally, serological and some molecular approaches cannot help to identify novel pathogenic agents and are, therefore, unsuitable for pathogen discovery [29].

Next Generation sequencing (NGS) technologies produce significantly more raw data than other molecular diagnostic assays, including Sanger sequencing, and are also capable of informing not just pathogen diagnostics but also epidemiology [30]. This is why whole genome sequencing of viral genomes by using new technologies plays an important role in the fight against emerging and re-emerging epidemics [31, 32]. The availability of high-throughput sequencing has also provided immense insights into the ecology of health care-associated pathogens [33]. Therefore, real-time sequencing of entire pathogen genomes has become a standard and indispensable research tool for the critical role of genomic surveillance in the prevention and control of emerging infectious diseases [34], which justifies why NGS can be considered a powerful strategy that also allows the discovery of novel potential viral pathogens [35, 36].

Considering pathogen surveillance in mind, bioinformatics tools and the combination of genomic and epidemiological data from viral infections can give essential information for understanding the past and the future of an epidemic, because genomic data generated by real-time sequencing can provide important information on how and when viruses were introduced in a particular site, their pattern and determinants of dissemination in



IntechOpen

neighboring locations and the extent of genetic diversity, i.e., its dynamics, making it possible to establish an effective surveillance framework on tracking the spread of infections to other geographic regions [23, 24, 36]. In this context, recently established international networks for real-time, portable genomic sequencing, genomic surveillance and data analysis made it possible to monitor the evolution of viral genomes, to understand the origins of outbreaks and epidemics, to predict future outbreaks and to assist in the maintenance of updated diagnostic methods [35, 36, 37]. Additionally, genomic surveillance framework allows to determine, through genome sequencing, the real-time molecular epidemiology of viruses circulating and co-circulating in different regions in a specific area, and also to detect and characterize the early emergence of new pathogens in large urban centers, generating data that can inform outbreak control responses [29, 36]. Generated data regarding the molecular, epidemiological, phylogenetic and geographical aspects of circulating viral pathogens in a specific setting contribute to a better understanding of those viral infections in a national and international context, assuming an important role in solving issues relevance to Public Health [37]. As a result, studies involving more in-depth molecular and dispersion analysis of circulating pathogens may help the World Health Organization appropriately adopt measures to control epidemics and to monitor the dynamics and spreading of new viral strains. However, even though NGS has advantages over diagnostics routine, all of the different strategies and technologies, developed by Illumina, Thermo Scientific, Oxford Nanopore and others, are not yet considered a panacea. Remaining challenges include dealing with high data throughput, which requires sophisticated computational processing as well as the annotation of large amounts of sequencing data, high DNA or RNA input sample requirements (in some cases hundreds of nanograms), which often raises the need for previous PCR-based amplification approaches. On top of all this, there are relatively



few researchers in the area with sufficient bioinformatics expertise and who are able to engage in near-patient or disease surveillance activities [37].

3. Bioinformatics tools and Phylogenetic tools

The advent of Next Generation Sequence (NGS) and advancements in bioinformatics present an opportunity to tap into new insights that are crucial to the establishment of an open, global digital surveillance system. NGS technologies have enabled the production and deposit of vast amounts of whole genomes into public repositories [38-40] ushering the field of genomics into era of big data. This has in turn increased the scale of genomic studies from the analysis of single or few genomes to an ever-increasing large number of genomes [41, 42].

Towards the development of global surveillance system, bioinformatics provides the tools to answer pertinent questions including the identification of organisms responsible for an outbreak, the source of an outbreak and evolutionary information of pathogens crucial for understanding the unique phenotypes such as drug resistance, virulence and disease outcome.

Several bioinformatic tools and pipelines have been developed to facilitate the processing, analysis and visualization of these data in order to derive useful information from it [43]. The major fields of interest addressed by these tools include comparative genomics which involves comparing the genetic content of one organism against that of another; prediction of the function of genes and sequences of the coding regions; identification of evolutionary events and inference of phylogenetic relationships. These fields of study play a critical role in elucidating pathogen evolution, niche adaptation, population structure and host-pathogen interaction. Furthermore, these findings inform vaccine and drug design, as well as the identification of virulence genes.



4. Bioinformatic Pipelines and Workflows

Bioinformatic pipelines and workflows comprise of a series of third-party executable command line software assembled to perform a specific task or analysis. A complete pipeline will, therefore, be able to support the end of analysis of a given field of study such as phylogenetics or variant detection. Pipelines can thus be broken down into two major components i.e. the data processing component and the analytical component that performs the core analysis of the pipeline. Below, we review some of the prominent bioinformatic pipelines and workflows that support the processing and analysis of NGS data to provide insights on relevant global surveillance of arboviral outbreaks.

5. Virus Discovery and Identification Tools

Viral discovery and identification from isolates and metagenomic samples present major challenges to bioinformatics in general. This is because viral genomes are prone to very high variability and deviation from reference genomes [44], continuous emergence of new viruses with no available references, high intrapopulation diversity, and the relative rareness of viral DNA fragments in metagenomic samples [45]. These challenges have largely been addressed through the following pipelines.

5.1 *Genome Detective*

Genome Detective (<http://www.genomedetective.com/app/>) is an easy to use web-based software application that assembles the genomes of viruses quickly and accurately, designed to generate and analyse whole or partial viral genomes directly from NGS reads within minutes [46]. The application gains accuracy by using a novel alignment method that uses a combination of both amino-acids and nucleotide scores to construct genomes by reference-based linking of de novo contigs. Speed and accuracy are also gained by using DIAMOND [47] with a UniProt90 reference dataset to sort viral taxonomy units. The use of DIAMOND and UniRef90



IntechOpen

allowed Genome Detective to identify viral short reads at least 1000 times faster than when Blastn and the viral nucleotide database of NCBI were used. The software was optimized using synthetic datasets to represent the great diversity of virus genomes. The application was then validated with next-generation sequencing data of hundreds of viruses.

5.2 VirusTAP: Viral Genome-Targeted Assembly Pipeline

One of the major difficulties in this process is the correct de novo assembly of viral genomes from crude metagenomic deep sequencing reads, including large amounts of bacteria and human related sequencing reads. Such read contaminations often force the server to overload during de novo assembly and might cause mis-assembly of the resultant contigs. Pre-filtering by host-mapping subtraction could lead to efficient de novo assembly, allowing the rapid and accurate procurement of a complete viral genome sequence. In addition to the accuracy of de novo assembly, the exclusion of human-related sequences can circumvent conflicting ethical issues by avoiding analyzing the personal genetic information of patients [48, 49].

VirusTAP is web-based, integrated NGS analysis tool designed to facilitate rapid and accurate viral genome assembly from raw reads by just clicking on several selections. Like genome detective, it ensures that non-viral reads are eliminated prior to de novo assembly in order to ensure performance is not compromised.

5.3 Virus Identification Pipeline (VIP)

VIP (<https://github.com/keylabivdc/VIP>) is a web-based virus discovery and identification tool [48]. With a single click, it will filter out background-related reads, classify reads on basis of nucleotide and remote amino acid homology, and perform phylogenetic analysis to provide evolutionary insights.



5.4 TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data

TAR-VIR is a non-reference based NGS analysis tool for the reconstruction of viral strains from metagenomic samples [48, 49]. It was developed to classify RNA viral reads from viral metagenomic data and also to produce the assembled viral strains (i.e. haplotypes) from classified reads. It mainly has two components: (1) Viral read classification using partial or remotely related reference genomes; (2) de novo assembly of viral haplotypes from recruited reads with PEHaplo [50, 51], which is a haplotype reconstruction tool. As TAR-VIR has a modular structure, the users have options to use other assembly tools after read classification in step (1).

6. Genotyping tools

While variant discovery and identification tools play a critical role in determining the pathogen responsible for the infection, they are unable to determine the subtype or quasispecies that is responsible for the outbreak. Arboviruses exist as a mixed population of genomic variants due to rapid replication and the error prone nature of viral RNA-dependent RNA polymerase (RdRp) [51]. Monitoring virus genotype diversity is therefore crucial to understand the emergence and spread of outbreaks. Genotyping tools provide an efficient workflow to enable researchers and public health practitioners to determine the strain that is responsible for the outbreak.

Most free-access bioinformatic programs used to classify the genetic profile of subtypes, genotypes, subgroups or groups of viruses are based on the use of similarity search tools to determine the genotype of a new sequence. These genotyping tools use a set of reference sequence genomes, carefully selected for the purpose of representing each individual genotype. The use of a number of reference sequences representing the genotype of a



given group increases the consistency and reproducibility of the data, thus ensuring a higher speed in the search for the data and offering greater and more complete information while ensuring that the results are not limited to an inadequate set of reference sequences that do not represent the information needed to identify the virus.

The similarity-based methods are useful for identifying recombination patterns in viral sequences, but they need further confirmation of their own phylogenetic methods and have no statistical support for their results.

Recently [52], four viral genotyping tools for yellow fever (YFV) (<https://www.genomedetective.com/app/typingtool/yellowfevervirus/>), dengue (DENV) (<https://www.genomedetective.com/app/typingtool/dengue/>), Chikungunya (CHIKV) (<https://www.genomedetective.com/app/typingtool/chikungunya/>) and Zika (ZIKV) (<https://www.genomedetective.com/app/typingtool/zika/>) were developed and linked to Genome Detective to enable phylogenetic classification below species level [53, 54].

6.1 Castor

The classification and annotation of virus genomes constitute important assets in the discovery of genomic variability, taxonomic characteristics and disease mechanisms. Existing classification methods are often designed for specific well-studied families of viruses [45]. Thus, the viral comparative genomic studies could benefit from more generic, fast and accurate tools for classifying and typing newly sequenced strains of diverse virus families.

CASTOR is a virus classification platform based on machine learning methods, inspired by a well-known technique in molecular biology: restriction fragment length polymorphism [55]. It simulates, in silico, the restriction digestion of genomic material by different enzymes into fragments. It uses two metrics to construct feature vectors for machine learning algorithms in the classification step. The performance of



IntechOpen

CASTOR, its genericity and robustness could permit the conduct of novel and accurate large-scale virus studies. The CASTOR web platform provides an open access, collaborative and reproducible machine learning classifiers. CASTOR can be accessed at (<http://castor.bioinfo.uqam.ca>).

7. Phylogenetic and Phylodynamic Tools

Phylogenetic tools are an extremely important resource used in the field of virology to study viral evolution, trace the origin of epidemics, establish the mode of transmission, investigate the occurrence of drug resistance or determine the origin of the virus in different body compartments. Thus, the tools developed by bioinformatics are fundamental to monitor the evolution of viral diversity, supporting studies of genomic sequence analysis, crucial for the surveillance of viral polymorphism, the development of new therapeutic strategies, the development of vaccine products or the appropriate choice products. Towards the development of a global surveillance outbreak surveillance system, the advances below have been made.

7.1 Nextstrain (<https://nextstrain.org/>)

Nextstrain is a real-time pathogen evolution tracking platform that implements cutting edge analysis and visualization of pathogen genome data [56]. It provides evolutionary information in the form of interactive visualizations to virologists, epidemiologists, public health officials and citizen scientists. It has been used to track various arboviral epidemics globally including West Nile Virus (WNV) in the Americas, Zika virus in 33 countries and Dengue virus outbreaks in 64 countries. The platform is continually updated with publicly available datasets to provide new insights into viral epidemic outbreaks globally in an intuitive and visually aesthetic manner.



8. Functional Prediction Tools

In disease surveillance, understanding the effect of mutations detected in the viral genomes through the methods identified above is invaluable in the development of relevant controls and interventions [57]. Many of these mutations serve as drug targets as well as provide insights into the response mechanism of the pathogens to existing interventions. A global surveillance system would therefore be incomplete without the capability to provide insights to the function of discovered mutations. Below we explore some of the tools that have been applied to understand the functional relevance of mutations found in arboviruses.

8.1 The SIFT (*Sorting Intolerant from Tolerant*)

The SIFT algorithm predicts the effect of coding variants on protein function [58, 59]. Since its introduction in 2001, SIFT has become one of the standard tools for characterizing missense variation. It has a corresponding website that provides users with predictions on their variants.

9. Conclusion

Augmenting epidemiological data with insights from genomic data provides a powerful tool for surveillance and control of disease outbreaks. Advances in bioinformatics particularly leverage large genomic datasets to determine pathogenic organisms responsible for the outbreak, the origin of the infection and mutations responsible unique phenotypic traits. This information is crucial for effective planning interventions and combating outbreaks. An area of research interest that remains to be explored is the development of online platforms to perform functional analyses of statistically significant mutations in arboviruses. This information is invaluable in the development of vaccines and identification of drug targets.



Acknowledgments

This work was supported by the ZiBRA2 project supported by the Brazilian Ministry of Health (SVS-MS) and the Pan American Organization (OPAS) and funded by Decit/SCTIE/MoH and CNPq (440685/2016-8 and 440856/2016-7); by CAPES (88887.130716/2016-00, 88881.130825/2016-00 and 88887.130823/2016-00). MG is supported by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ.

Conflict of Interest

The authors declare no conflict of interest.

Appendices and Nomenclature

RT-qPCR: Real Time quantitative Polymerase chain reaction

NGS: Next Generation Sequencing

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

VIP: Virus Identification Pipeline

TAR-VIR: Targeted Viral

RdRp: RNA-dependent RNA polymerase

YFV: Yellow Fever virus

DENV: Dengue virus

CHIKV: Chikungunya virus

ZIKV: Zika virus

WNV: West Nile Virus

SIFT: Sorting Intolerant from Tolerant



References

- [1] WHO. Vector-borne diseases. 2017. Available from: <https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases>.
- [2] World Economic Forum. Global Risks 2015. World economic forum. Insight Report. 10th Edition. 2015. Available from: http://www3.weforum.org/docs/WEF_Global_Risks_2015_Report15.Pdf. Pdf. 2015.
- [3] LaBeaud AD. Mint: Why arboviruses can be neglected tropical diseases. *PLoS Negl Trop Dis*. 2008; 25:6-247. DOI: 10.1371/journal.pntd.0000247.
- [4] Powell JR. Mint: Mosquito-borne human viral diseases: Why *Aedes aegypti*? *The American journal of tropical medicine and hygiene*. 2018; 98:1563-5. DOI: 10.4269/ajtmh.17-0866.
- [5] Leta S, Beyene TJ, De Clercq EM, Amenu K, Kraemer MU, Revie CW. Mint: Global risk mapping for major diseases transmitted by *Aedes aegypti* and *Aedes albopictus*. *International Journal of Infectious Diseases*. 2018; 67:25-35. DOI: 10.1016/j.ijid.2017.11.026.
- [6] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF. Mint: The global distribution and burden of dengue. *Nature*. 2013; 496:504-544. DOI: 10.1038/nature12060.
- [7] PAHO. 2019a. Dengue and Severe Dengue, Cases and Deaths for subregions of the Americas. 2019. Available from: <http://www.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-regional-en/261-dengue-reg-ano-en.html>
- [8] PAHO. 2019b. Chikungunya total cases. Available from: Available from: <http://www.paho.org/data/index.php/en/mnu-topics/chikv-en/551-chikv-subregions-en.html>
- [9] PAHO. 2019c. Zika total cases. 2019. Available from: <http://www.paho.org/data/index.php/en/mnu-topics/zika.html>
- [10] WHO. 2019a. Neglected Tropical Diseases in The Eastern Mediterranean Region. 2019. Available from: https://apps.who.int/iris/bitstream/handle/10665/275463/Fact_Sheet_CDT_2018_EN_20491.pdf?ua=1
- [11] WHO. 2019b. Dengue and severe dengue. 2019. Available from: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>



- [12] WHO. Emergencies preparedness, response. Chikungunya. 2018. Available from: <https://www.who.int/csr/don/archive/disease/chikungunya/en/>
- [13] LaBeaud AD. Mint: Why arboviruses can be neglected tropical diseases. *PLoS neglected tropical diseases*. 2008; 25:6-247. DOI: 10.1371/journal.pntd.0000247.
- [14] Schar DL, Yamey GM, Machalaba CC, Karesh WB. Mint: A framework for stimulating economic investments to prevent emerging diseases. *Bulletin of the World Health Organization*. 2018; 96-138. DOI: 10.2471/BLT.17.199547.
- [15] Messina JP, Brady OJ, Golding N, Kraemer MU, Wint GW, Ray SE, Pigott DM, Shearer FM, Johnson K, Earl L, Marczak LB. Mint: The current and future global distribution and population at risk of dengue. *Nature microbiology*. 2019; 01:10-11. DOI: /10.1038/s41564-019-0476-8.
- [16] Franklins LH, Jones KE, Redding DW, Abubakar I. Mint: The effect of global change on mosquito-borne disease. *The Lancet Infectious Diseases*. 2019; 01:18-124. DOI: 10.1525/abt.2017.79.3.169.
- [17] Zanotto, P.M.A. and L.C.C. Leite, Mint: The Challenges Imposed by Dengue, Zika, and Chikungunya to Brazil. *Front Immunol*, 2018; 9:1960-1964. DOI: 10.3389/fimmu.2018.01964.
- [18] Machado AA, Estevan AO, Sales A, da Silva Brabes KC, Croda J, Negrão FJ. Mint: Direct costs of dengue hospitalization in Brazil: public and private health care systems and use of WHO guidelines. *PLoS neglected tropical diseases*. 2014; 4:8-104. DOI: 10.1371/journal.pntd.0003104.
- [19] Bangert M, Latheef AT, Pant SD, Ahmed IN, Saleem S, Rafeeq FN, Abdulla M, Shamah F, Mohamed AJ, Fitzpatrick C, Velayudhan R. Mint: Economic analysis of dengue prevention and case management in the Maldives. *PLoS neglected tropical diseases*. 2018; 27:12-96. DOI: 10.1371/journal.pntd.0006796.
- [20] Staples JE, Shankar MB, Sejvar JJ, Meltzer MI, Fischer M. Mint: Initial and long-term costs of patients hospitalized with West Nile virus disease. *The American journal of tropical medicine and hygiene*. 2014; 3:402-9. DOI: 10.4269/ajtmh.13-0206.
- [21] Rossi G, Karki S, Smith RL, Brown WM, Ruiz MO. Mint: The spread of mosquito-borne viruses in modern times: A spatio-temporal analysis of dengue and chikungunya. *Spatial and spatio-temporal epidemiology*. 2018; 26:113-25. DOI: 10.1016/j.sste.2018.06.002.
- [22] Messina JP, Brady OJ, Golding N, Kraemer MU, Wint GW, Ray SE, Pigott DM, Shearer FM, Johnson K, Earl L, Marczak LB. Mint: The



current and future global distribution and population at risk of dengue. *Nature microbiology*. 2019; 01:10-111. DOI: 10.1038/s41564-019-0476-8.

[23] Gardy, J. L.; Loman, N. J. Mint: Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*. 2017; 1:2-256. DOI: 10.1038/nrg.2017.88.

[24] Grubaugh, N. D. Mint: Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*. 2019; 4:10-19. DOI: 10.1038/s41564-018-0296-2.

[25] D.M. Morens, G.K. Folkers, A.S. Fauci. Mint: The challenge of emerging and re-emerging infectious diseases. *Nature*. 2004; 430:242–249. DOI: 10.1038/nature02759.

[26] Daszak P, Cunningham A.A, Hyatt A.D. Mint: Emerging infectious diseases of wildlife—threats to biodiversity and human health. *Science*. 2000; 287:443–449. DOI: 10.1126/science.287.5452.443.

[27] Morse S.S. Mint: Factors in the emergence of infectious diseases. *Emerg. Infect. Dis*. 1995; 1:7–15. DOI: 10.3201/eid0101.950102.

[28] Versalovic J, Lupski J.R. Mint: Molecular detection and genotyping of pathogens: more accurate and rapid answers. *Trends Microbiol*. 2002; 10:15-21. DOI: 12377563.

[29] Sabat A.J, Budimir A, Nashev D, Sá-Leão R, van Dijk J.M, Laurent F. Mint: Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill*. 2013; 18 :20-380. DOI: 10.2807/ese.18.04.20380-en.

[30] Shendure J, Ji H. Mint: Next-generation DNA sequencing. *Nat. Biotechnol*. 2008; 26:1135- 45. DOI: 10.1038/nbt1486.

[31] Haagmans, B.L., Andeweg, A.C., and Osterhaus, A.D.M.E. Mint: The application of genomics to emerging zoonotic viral diseases. *PLoS Pathog*. 2009; 5:100-557. DOI: 10.1371/journal.ppat.1000557.

[32] McHardy, A.C., and Adams, B. Mint: The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathog*. 2009; 5:10-56. DOI: 10.1371/journal.ppat.1000566.

[33] Tang P, Gardy J.L. Mint: Stopping outbreaks with real-time genomic epidemiology. *Genome Med*. 2014; 6:1-104. DOI: 10.1186/s13073-014-0104-4.

[34] Holmes, E.C. Mint: Viral evolution in the genomic age. *PLoS. Biol*. 2007; 5:2-78. DOI: 10.1371/journal.pbio.0050278.

[35] Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K. Mint: Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature protocols*. 2017; 12:12-61. DOI: 10.1038/nprot.2017.066.



[36] Thézé, J., Li, T., du Plessis, L., Bouquet, J., Kraemer, M. U., Somasekar, S. Mint: Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. *Cell host & microbe*. 2018; 23:855–864. DOI: 10.1016/j.chom.2018.04.017.

[37] Loman N.J, Constantinidou C, Chan J.Z.M, Halachev M, Sergeant M, Penn C.W. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol*. 2012. 10:599-606. DOI: 10.1038/nrmicro2850.

[38] Nader I. AL-Dewik and M. Mint: Walid Qoronfleh, “Genomics and Precision Medicine: Molecular Diagnostics Innovations Shaping the Future of Healthcare in Qatar,” *Advances in Public Health*. 2019; 2-44-76. DOI: <https://doi.org/10.1155/2019/3807032>.

[39] Zhang J, Chiodini R, Badr A, Zhang G. Mint: The impact of next-generation sequencing on genomics. *J Genet Genomics*. 2011; 38:95–109. DOI: 10.1016/j.jgg.2011.02.003.

[40] Koboldt, Daniel C., Karyn M. Steinberg, David E. Larson, Richard K. Wilson and E. R. Mardis. Mint: The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*. 2013; 155: 27-38. DOI: 10.1016/j.cell.2013.09.006.

[41] Elliott, L. T., K. Sharp, F. Alfaro-Almagro, S. Shi, K. L. Miller, G. Douaud, J. Marchini and S. M. Smith. Mint: Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*. 2018; 562: 210-216. DOI: 10.1038/s41586-018-0571-7.

[42] Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., & Beyene, J. Mint: Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics: HGP*. 2009; 86:90-93. DOI:10.4061/2009/869093.

[43] Hwang, B., J. H. Lee and D. Bang. Mint: Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*. 2018; 50:8-96. DOI: 10.1038/s12276-018-0071-8.

[44] Manso CF, Bibby DF, Mbisa JL. Mint: Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. *Sci Rep*. 2017; 7:41-73. DOI: 10.1038/s41598-017-02239-5.

[45] Rose R, Constantinides B, Tapinos A et al. Mint: Challenges in the analysis of viral metagenomes. *Virus Evol*. 2016; 2:01-22. DOI: 10.1093/ve/vew022.

[46] Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, Alcantara LC, Vanden Eynden E, Vandamme AM, Deforche K, de Oliveira T. Mint: Genome Detective: an automated system for virus



identification from high-throughput sequencing data. *Bioinformatics*. 2018; 2:23-98. DOI: 10.1093/bioinformatics/bty695.

[47] B. Buchfink, C. Xie, and D. H. Huson. Mint: Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*. 2014 12:20-59. DOI: 10.1038/nmeth.3176.

[48] Amine M. Remita, Ahmed Halioui, Abou Abdallah Malick Diouara, Bruno Daigle, Golrokh Kiani and Abdoulaye Banire Diallo. Mint: A machine learning approach for viral genome classification. *BMC Bioinformatics*. 2017. 18:2-08. DOI: 10.1186/s12859-017-1602-3.

[49] Chen J, Huang J, Sun Y. Mint: TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. *BMC Bioinformatics*. 2019; 20:3-05. DOI: 10.1186/s12859-019-2878-2.

[50] Chen J, Zhao Y, Sun Y. Mint: De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding. *Hancock J. Bioinformatics* 2018; 34:2927–35. DOI: 10.1093/bioinformatics/bty202.

[51] Chen J, Huang J, Sun Y. Mint: TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. *BMC Bioinformatics*. 2019; 20:3-05. DOI: 10.1186/s12859-019-2878-2.

[52] Faria, N. R. Mint: Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*. 2018; 36: 894-899. DOI: 10.1126/science.aat7115.

[53] Fonseca V, Libin PJK, Theys K. Mint: A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes. Rodriguez-Barraquer I (ed.). *PLoS Negl Trop Dis*. 2019; 13:7-231. DOI: 10.1371/journal.pntd.0007231.

[54] M. A. Remita, A. Halioui, A. A. Malick Diouara, B. Daigle, G. Kiani, and A. B. Diallo. Mint: A machine learning approach for viral genome classification. *BMC Bioinformatics*. 18; 1:217-208. DOI: 10.1186/s12859-017-1602-3.

[55] Hadfield J, Megill C, Bell SM. Mint: Nextstrain: real-time tracking of pathogen evolution. Kelso J (ed.). *Bioinformatics* 2018; 34:4121–3. DOI: 10.1093/bioinformatics/bty407.

[56] Alexander TC, Laura DK. Mint: Insights into Arbovirus Evolution and Adaptation from Experimental Studies. *Viruses*. 2010; 12: 2594–2617. DOI: 10.3390/v2122594.

[57] Chen J, Huang J, Sun Y. Mint: TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. *BMC Bioinformatics*. 2019; 20:3-65. DOI: 10.1186/s12859-019-2878-2.



IntechOpen

[58] Li, Y. Mint: VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci. Rep.* 2016; 6:23-774. DOI: 10.1038/srep23774.

[59] Sim N-L, Kumar P, Hu J. Mint: SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 2012; 40:452-7. DOI: 10.1093/nar/gks539.



1.3 Justification

Despite the growing knowledge about circulating and co-circulating arboviruses in Brazil, many questions remain unanswered about their vectors and reservoirs, their pathogenesis, their genetic diversity, and the potential synergistic effects of infection or co-infection between them or with other circulating viruses. These questions highlight the need for research, especially through a network, to optimize surveillance, patient follow-up, and public health intervention during epidemics. Given the rapid spread of ZIKV and CHIKV across Central and South America in recent years the potential for neurological complications as well as the lack of effective diagnosis, vaccine and therapy, these arboviruses are viewed as a major public health issue across the Americas. With the growing volume of data generated by infections caused by these viruses and research carried out on the human genome, other species and pathogens, the need for highly efficient computer systems to assist in processing this volume of information is even more accentuated.

Phylogenetic tools are significant resources used in virology to study viral evolution, trace the origin of epidemics, establish the transmission mode, research the occurrence of drug resistance, or determine the source of the virus in different body compartments. Therefore, the tools developed by bioinformatics are essential to monitor the evolution of viral diversity and support genomic sequence analysis studies. They are also crucial for the surveillance of viral polymorphism in the development of new therapeutic strategies and vaccines.

All open access bioinformatics programs used to classify the genetic profile of virus subtypes, genotypes, subgroups, or groups are based on similarity search tools to determine the genotype of a new sequence. Similarity-based methods help identify recombination patterns in viral sequences, but they require further confirmation of their phylogenetic methods and lack statistical support for their results.

Automated genotyping tools use a set of reference sequence genomes, carefully selected to represent each genotype. The use of several reference sequences that represent the genotype of a given group increases the consistency and reproducibility of the data. This accelerates the search and improves the quality of the information by excluding results with an inadequate set of reference sequences.

1.4 Main objective

To develop user-friendly, web-based computational tools and underlying methods to analyze raw data from next-generation sequencing (NGS) platforms and perform viral genomic characterization to identify viral infectious agents in clinical samples from infected individuals.

1.4.1 Specific objectives

- To develop a tool for assembling viral genomes using NGS platforms for short and long reads.
- To develop tools for the genomic characterization and genotypes/lineages assignment of emerging and reemerging, viral pathogens.
- To make available in the tool phylogenetic trees, based on bootstrap analysis of the monophyletic groups of sequences submitted by the user.
- To make assembly information available in the tool, such as the method used, alignment, mutations, contigs, etc.

1.5 Ethical approval

The availability of these samples for research purposes during outbreaks of national concern is allowed according to the terms of the 510/2016 Resolution of the National Ethical Committee for Research – Brazilian Ministry of Health (CONEP - Comissão Nacional de Ética em Pesquisa, Ministério da Saúde). It authorizes, without the necessity of an informed consent, the use of clinical samples collected in the Brazilian Central Public Health Laboratories to accelerate knowledge building and contribute to surveillance and outbreak response.

CHAPTER 2: GENOME DETECTIVE: AN AUTOMATED SYSTEM FOR VIRUS IDENTIFICATION FROM HIGH-THROUGHPUT SEQUENCING DATA

This Chapter presents the quick and efficient development of viral genome analyzers that are challenging due to their high variability and deviation from reference genomes. This is compounded by the increased speed of identification, the continuous emergence of new viruses, and the relative recessiveness of viral fragments in metagenomic analyses. The tool called Genome Detective was developed to solve this problem. This tool was designed and thought to allow fast assembly and real-time analysis of partial and complete viral genomes directly from Next Generation Sequencing (NGS) readings in just a few minutes, in addition to allowing the discovery of new viruses, through assembly “de novo” from samples originally submitted to metagenomics experiments.

The Chapter is presented in the form of a research article. It has been published in the journal, *Bioinformatics*. The published manuscript is hereby presented in the journal format.

Manuscript Published: Michael Vilsker, Yumna Moosa, Sam Nooij, Vagner Fonseca, Yoika Ghysens, Korneel Dumon, Raf Pauwels, Luiz Carlos Alcantara, Ewout Vanden Eynden, Anne-Mieke Vandamme, Koen Deforche, Tulio de Oliveira, Genome Detective: an automated system for virus identification from high-throughput sequencing data, Bioinformatics, Volume 35, Issue 5, 01 March 2019, Pages 871–873, <https://doi.org/10.1093/bioinformatics/bty695>

Sequence analysis

Genome Detective: an automated system for virus identification from high-throughput sequencing data

Michael Vilsker¹, Yumna Moosa², Sam Nooij³, Vagner Fonseca^{2,4,5}, Yoika Ghysens¹, Korneel Dumon¹, Raf Pauwels¹, Luiz Carlos Alcantara^{4,5,6}, Ewout Vanden Eynden⁷, Anne-Mieke Vandamme^{7,8}, Koen Deforche^{1,*} and Tulio de Oliveira^{1,2,*}

¹Emweb bvba, 3020 Herent, Belgium, ²KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, Nelson R Mandela School of Medicine, College of Health Sciences, University of KwaZulu-Natal, Durban 4001, South Africa, ³The Dutch National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands, ⁴Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, ⁵Laboratory of Hematology Genetic and computational Biology, Gonçalo Moniz Research Center, Oswaldo Cruz Foundation (LHGB/CPqGM/FIOCRUZ), Bahia, Brazil, ⁶Laboratório de Flavivirus, IOC, Fundação Oswaldo Cruz, ⁷KU Leuven, Department of Microbiology and Immunology, Rega Institute for Medical Research, Clinical and Epidemiological Virology, Leuven, Belgium, and ⁸Center for Global Health and Tropical Medicine, Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 4, 2018; revised on July 9, 2018; editorial decision on July 30, 2018; accepted on August 14, 2018

Abstract

Summary: Genome Detective is an easy to use web-based software application that assembles the genomes of viruses quickly and accurately. The application uses a novel alignment method that constructs genomes by reference-based linking of *de novo* contigs by combining amino-acids and nucleotide scores. The software was optimized using synthetic datasets to represent the great diversity of virus genomes. The application was then validated with next generation sequencing data of hundreds of viruses. User time is minimal and it is limited to the time required to upload the data.

Availability and implementation: Available online: http://www.genomedetective.com/app/typing_tool/virus/.

Contact: koen@emweb.be or deoliveira@ukzn.ac.za

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the rapidly expanding field of genomics, our ability to produce data far exceeds our capacity to analyze and extract meaningful information. Analysis of viral data is particularly challenging given the high variability of viruses and their deviation from reference genomes, the increasing speed of identification, the continuous

emergence of new viruses and the relative scarcity of viral fragments in metagenomic samples (Rose *et al.*, 2016).

The quality of available tools varies and most require specialized computing skills and access to powerful hardware in order to analyze next generation sequencing (NGS) data and/or high-throughput Sanger data. In response to this need, we have developed Genome Detective, a web-based bioinformatics pipeline to accurately and

Table 1. Validation datasets

Publication	PMID	Description	Number of datasets	Expected number of viruses	Assigned number of viruses	Average reconstructed genome size (%)	Number of additional viruses
1	26 559 140	Synthetic virome	8	64	57	92	9
2	Pending (bioRxiv)	Single virus—HIV	14	13	13	93	1
Unpublished	PRJNA434 385 (SRA)	Single virus—HIV	94	94	94	95	15
3	25 609 811	Single virus—RSV	12	12	12	98	1
4	25 056 894	Single virus—norovirus	12	12	12	99	7
5	26 071 329	Single virus—influenza	10	10	10 (80 segments)	94	26
6	24 055 451	Single virus—MERS	14	14	14	94	0
7	28 748 110	Metagenomic—pig fecal	20	20	20 (220 segments)	90	143
8	24 695 106	Metagenomic—human fecal	20	66	25	83	35
	—	—	204	305	257	—	237

Note: For the validation of Genome Detective (GD) we used 204 datasets from seven studies. This table lists the PMID of the publications, a description of the data, number of datasets, number of viruses originally identified, number of viruses for which GD reconstructed whole genomes (i.e. >80% of the whole genome and high NT/AA score) and number of viruses that GD additionally detected (i.e. <80% of the whole genome or low NT/AA score). Detailed information such as (SRA files list and full results are seen in Supplementary Material).

quickly identify, assemble and classify all known viruses present in NGS and Sanger sequencing data.

2 Systems and methods

Genome Detective accepts unprocessed paired-end or single reads generated by NGS platforms in FASTQ format and/or processed FASTA sequences. For FASTQ files, low-quality reads are filtered and adapters trimmed with Trimmomatic (Bolger et al., 2014). The quality of the reads is visualized using FastQC (Brown et al., 2017) before and after trimming. Candidate viral reads are identified using the protein-based alignment method DIAMOND (Buchfink et al., 2015). We used the viral subset of the Swissprot UniRef90 protein database, which contains representative clusters of proteins linked to taxonomy IDs, to improve sensitivity and speed. The Swissprot UniRef90 is constantly updated and at the time of the submission of this paper, the viral subset of this database contained 494 134 protein clusters. At the same time, also the NCBI RefSeq database is constantly updated, and at the time of the submission of this paper, the viral subset of this database contained 7560 unique taxonomic IDs. Genome Detective has an automated procedure to download new versions of the reference databases and the current version and the number of viral taxonomy IDs identified are shown on the interface.

The speed and accuracy of Genome Detective was also improved by first sorting short reads into groups, or buckets. Our objective was to run a separate metagenomic *de novo* assembly in each bucket, so all reads of one virus species needed to be assigned to the same bucket. Each bucket is then identified using the taxonomy ID of the lowest common ancestor (LCA) of the hits identified by DIAMOND. However, some reads that represented the same viral species were assigned to buckets at different taxonomic ranks. We solved this problem by either distributing the reads from the node downwards, or collapsing them upwards, by comparing the number of reads identified at each node of the taxonomy tree versus in all descendant nodes. In addition, given that metagenomic studies are accelerating (reviewed in Rose et al., 2016), an increasing number of reference sequences are of novel viruses that have not yet been classified. This causes the LCA taxonomy ID to be unspecific for a number of Uniref clusters, and in the analysis of hits identified by DIAMOND. To avoid these problems, while retaining the sequence themselves, we excluded the taxonomic classification of these viruses in LCA algorithms.

Once all of the reads have been sorted in buckets, each bucket is then *de novo* assembled separately using SPAdes (Bankevich et al., 2012) for single-ended reads or metaSPAdes (Bankevich et al., 2012) for paired-end reads. Blastx and Blastn are used to search for candidate reference sequences against the NCBI RefSeq virus database. Genome Detective combines the results for every detected contig at amino acid and nucleotide (nt) level with by calculating a total score that is a sum of the total nt score plus total amino acid score. We then chose the five best scoring references for each contig to be used during the alignment.

The contigs for each individual species are joined using Advanced Genome Aligner (AGA) (Deforche, 2017), which is a new dynamic programming algorithm. AGA is designed to compute the optimal global alignment considering simultaneously the alignment of all annotated coding sequences of a reference genome. AGA builds further on the optimal alignment algorithms first proposed by Needleman–Wunsch (Smith and Waterman, 1981), Smith–Waterman (Smith and Waterman, 1981) and Gotoh (Gotoh, 1982), by expanding the induction state with additional state parameters. This makes alignments using AGA, and therefore Genome Detective, more sensitive and accurate as both nt and protein scores are taken into account in order to produce a consensus sequence from the *de novo* contigs.

A report is generated, referring to the final contigs and consensus sequences, available as FASTA files. The report also contains detailed information on filtering, assemblage and consensus sequence. Web-based (using the JWt libraries) graphics are available for viral species, genome images, alignment viewer, nt and amino acid similarity measures and read counts. In addition, the user can produce a bam file with BWA (Li and Durbin, 2009) using the reference or *de novo* consensus sequence by selecting the detailed report (Supplementary Fig. S1) and access viral phylogenetic identification tools (de Oliveira et al., 2005) directly from the interface.

3 Testing and validation

We first validated Genome Detective using a synthetic virus dataset (NCBI SRA: SRR3458562-SRR3458569), originally prepared to optimize laboratory-based virus extraction procedures, in which viruses were carefully selected to cover the range of naturally occurring diversity (Conceição-Neto et al., 2015). This published dataset also includes carefully validated quantitative results, confirmed with quantitative PCR. Genome Detective identified all of the viruses in

the synthetic dataset. We then validated Genome Detective with real clinical datasets. In total, we analyzed 208 datasets, which are available via Sequence Read Archive (SRA) or the European Nucleotide Archive. We then compared our results to the published results and found a >95% concordance, successfully identifying 257 viral species (Table 1 and Supplementary Table). These included single viruses with unsegmented (HIV) and segmented genomes (Influenza A, Rotavirus, MERS) from amplicon-based NGS sequenced as well as unbiased metagenomic datasets (Table 1). Overall, precision, sensitivity and specificity were high, with the exception of 20 metagenomic datasets from human fecal (ERR233412-ERR233431), which had scarce viral reads (Supplementary Table).

We compared our assignment results with IVA (Hunt *et al.*, 2015) and with drVM (Lin and Liao, 2017), which is a new and accurate method for efficient genome assembly of viruses. When the HIV-1 runs were compared with IVA, our web-based application reduced the processing time needed for assembling whole viral genomes by a factor of 10 (10–500-fold) and provided longer and more accurate contigs. In order to compare our results with drVM, we used five datasets (SRR1170797, SRR1106548, DRR049387, SRR062073 and ERR690519). These were the same datasets that the authors of drVM used to compare with three other similar tools, SURPI (Naccache *et al.*, 2014), VIP (Li *et al.*, 2016) and VirusTap (Yamashita *et al.*, 2016). We found that, in general, Genome Detective creates longer, more accurate contigs than drVM, SURPI, VIP and VirusTap. In addition, Genome Detective speed is similar or faster than the four other mentioned tools (Supplementary Material). For example, we assembled a near complete genome (length 8.334 bp) of HIV-1 (SRR1106548) in 430 s, whereas VirusTap identified a 2.896 bp contig in 1.388 s and drVM identified a 3.005 bp segment in 608 s. For the Rotavirus reads (DRR049387), Genome Detective identified all of the 11 segments of Rotavirus A (segment 1–11) in one contig, each covering 97–100% of each segment, whereas drVM identified only 7 segments from 13 contigs. The time for this run in Genome Detective was 440 versus 464 s of drVM (Lin and Liao, 2017). For Influenza A virus (ERR690519), we identified the same eight segments as drVM in less than half of the time (Supplementary Table S3).

4 Discussion

Genome Detective was developed to generate and analyze whole or partial viral genomes directly from NGS reads within minutes. Speed and accuracy were gained by using DIAMOND with a UniProt90 reference dataset to sort viral taxonomy units. The use of DIAMOND and UniRef90 allowed Genome Detective to identify viral short reads at least 1000 times faster than if we used Blastn and the viral nt database of NCBI (Buchfink *et al.*, 2015). Accuracy was also gained by joining contigs with a novel alignment method that uses amino acids and nt scores to create *de novo* contigs. Despite the use of only RefSeq for the identification of virus species, sensitivity and specificity were maintained due to the use of both nt and amino acid similarity scores. We found that for large NGS and metagenomic datasets, Virus Detective substantially reduces computational cost without compromising the quality of the result. However, the construction of *de novo* whole genomes from metagenomic samples depends on the number of reads, the virus genome size and read

length. Our pipeline also allows detailed displays of data and results. Furthermore, Genome Detective is linked to our popular virus-specific typing tools (>3 million submissions, de Oliveira *et al.*, 2005), which allow phylogenetic classification below species level. User time is minimal; it is limited to the time required to upload the data. In conclusion, Genome Detective is a web-based pipeline that allows raw NGS data to be assembled into *de novo* complete viral genomes in a fast and accurate manner.

Funding

Supported by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI), a Royal Society Newton Advanced Fellowship (TdO), the VIROGENESIS project receives funding from the European Union's Horizon 2020 Research and Innovation Program (under Grant Agreement no. 634650) and the National Institutes of Health Common Fund, grant number U24HG006941. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to acknowledge the contribution of Annelies Kroneman, Harry Vennema, Roel Standaert, Pieter Libin and Kristof Theys.

Conflict of Interest: none declared.

References

- Bankevich, A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
- Bolger, A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Brown, J. *et al.* (2017) FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, **33**, 3137–3139.
- Buchfink, B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
- Conceição-Neto, N. *et al.* (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific Reports*, **5**, 16532. Retrieved from <http://dx.doi.org/10.1038/srep16532>.
- de Oliveira, T. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.
- Deforche, K. (2017) An alignment method for nucleic acid sequences against annotated genomes. doi.org/10.1101/200394.
- Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
- Hunt, M. *et al.* (2015) IVA: accurate *de novo* assembly of RNA virus genomes. *Bioinformatics*, **31**, 2374–2376.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Li, Y. *et al.* (2016) VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci. Rep.*, **6**, 23774.
- Lin, F.H. and Liao, Y.C. (2017) drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. *Gigascience*, **6**, 1–10.
- Naccache, S.N. *et al.* (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.*, **24**, 1180–1192.
- Rose, R. *et al.* (2016) Challenges in the analysis of viral metagenomes. *Virus Evol.*, **2**, vew022.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Yamashita, A. *et al.* (2016) VirusTAP: viral genome-targeted assembly pipeline. *Front. Microbiol.*, **7**, 32. [PMCID: 4751111] [PubMed]

CHAPTER 3: A COMPUTATIONAL METHOD FOR THE IDENTIFICATION OF DENGUE, ZIKA AND CHIKUNGUNYA VIRUS SPECIES AND GENOTYPES

This Chapter presents the quick and efficient development of three genotyping tools for DENV, ZIKV, CHIKV. The tools method of the tools is to genotype by phylogenetic inference for this they do the aligning the user-submitted sequence with a carefully selected set of predefined reference strains, followed by Neighbour Joining (NJ) phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequences is assigned the genotype of the reference strain with the highest bootstrap scores (>70%). The tools allow high-throughput classification of these virus species and genotypes in seconds by providing users with a classified genotype report along with phylogenetic trees of the submitted sequences.

The Chapter is presented in the form of a research article. It has been published in the journal, *PLOS Neglected Tropical Diseases*. The published manuscript is hereby presented in the journal format.

Manuscript Published: Fonseca V, Libin PJK, Theys K, Faria NR, Nunes MRT, et al. (2019) A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes. PLOS Neglected Tropical Diseases 13(5): e0007231.

<https://doi.org/10.1371/journal.pntd.0007231>

RESEARCH ARTICLE

A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes

Vagner Fonseca^{1,2,3*}, Pieter J. K. Libin^{4,5*}, Kristof Theys^{5*}, Nuno R. Faria⁶, Marcio R. T. Nunes⁷, Maria I. Restovic⁸, Murilo Freire⁸, Marta Giovanetti¹, Lize Cuypers⁵, Ann Nowé⁴, Ana Abecasis⁹, Koen Deforche¹⁰, Gilberto A. Santiago¹¹, Isadora C. de Siqueira⁸, Emmanuel J. San², Kaliane C. B. Machado⁸, Vasco Azevedo³, Ana Maria Bispo-de Filippis¹, Rivaldo Venâncio da Cunha¹², Oliver G. Pybus⁶, Anne-Mieke Vandamme^{5,9}, Luiz C. J. Alcantara^{1,3*}, Tulio de Oliveira^{2*}



OPEN ACCESS

Citation: Fonseca V, Libin PJK, Theys K, Faria NR, Nunes MRT, Restovic MI, et al. (2019) A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes. *PLoS Negl Trop Dis* 13(5): e0007231. <https://doi.org/10.1371/journal.pntd.0007231>

Editor: Isabel Rodríguez-Barraquer, University of California San Francisco, UNITED STATES

Received: August 14, 2018

Accepted: February 11, 2019

Published: May 8, 2019

Copyright: © 2019 Fonseca et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant information of the data is within the paper and details were included as Supporting Information files.

Funding: Supported by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI), a Royal Society Newton Advanced Fellowship (TdO), the VIROGENESIS project receives funding from the European Union's Horizon 2020 Research and Innovation Programme (under Grant Agreement

1 Laboratório de Flavivírus, IOC, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, **2** KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZuluNatal, Durban, South Africa, **3** Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **4** Artificial Intelligence Lab, Department of Computer Science, Vrije Universiteit Brussel, 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** Department of Zoology, University of Oxford, Oxford, United Kingdom, **7** Evandro Chagas Institute, Ministry of Health, Ananindeua, Brazil, **8** Laboratório de Patologia Experimental, Fundação Oswaldo Cruz, Salvador, Brazil, **9** Center for Global Health and Tropical Medicine, Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal, **10** EMWEB (private company), Herent, Belgium, **11** Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, United states of America, **12** Coordenação de Vigilância em Saúde e Laboratórios de Referências, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

* These authors contributed equally to this work.

* alcantaraluz42@gmail.com (LCJA); tuliodna@gmail.com (TDO)

Abstract

In recent years, an increasing number of outbreaks of Dengue, Chikungunya and Zika viruses have been reported in Asia and the Americas. Monitoring virus genotype diversity is crucial to understand the emergence and spread of outbreaks, both aspects that are vital to develop effective prevention and treatment strategies. Hence, we developed an efficient method to classify virus sequences with respect to their species and sub-species (i.e. serotype and/or genotype). This tool provides an easy-to-use software implementation of this new method and was validated on a large dataset assessing the classification performance with respect to whole-genome sequences and partial-genome sequences. Available online: <http://krisp.org.za/tools.php>.

Author summary

Dengue (DENV), Chikungunya (CHIKV) and Zika (ZIKV) are considered major public health challenges. In addition to the epidemic caused by DENV, which has been described in many tropical countries, the introduction of CHIKV and ZIKV in these countries is a major public health concern. These arboviruses are primarily transmitted by mosquitoes of the species *Ae. Aegypti* and its related diseases result in increased financial costs associated with

no. 634650) and the National Institutes of Health Common Fund, grant number U24HG006941. Pieter Libin was supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek - Vlaanderen). This work was supported by Decit/SCTIE/MoH and CNPq (440685/2016-8 and 440856/2016-7); by CAPES (88887.130716/2016-00, 88887.130825/2016-00 and 88887.130823/2016-00); and by EU's Horizon 2020 Programme through ZIKAlliance (PRES-005-FEX-17-4-2-33). Ana Abecasis was supported by Fundação para a Ciência e Tecnologia (FCT) through funds to GHTM-UID/Multi/04413/2013. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Dr. Koen Deforche is one of the owners of the commercial company, EMWEB.

diagnosis and treatment. To support the design of efficient diagnosis, prevention and treatment strategies, a bioinformatics tool has been developed for the genotyping of these viruses based on appropriate evolutionary models in an automatic, accurate and rapid manner. A set of virus reference sequences was obtained from GenBank and used for the development of the tools. This process involved the alignment of the reference sequences followed by phylogenetic tree reconstructions. To assign the genotypes uploaded by the user, the tool analyses the sequences one by one, genotypes through identification, alignment and phylogenetic reconstruction. This computational method allows the high-throughput classification of these virus species and genotypes in seconds. As shown experimentally, genotypes are classified most confidently using the envelope gene or complete genome sequences.

Introduction

In the recent years, an increasing number of outbreaks of Dengue (DENV), Chikungunya (CHIKV) and Zika (ZIKV) viruses have been reported in Asia and the Americas [1–3]. The predominant mosquito species transmitting DENV, CHIKV and ZIKV, are *Aedes aegypti* and *Aedes Albopictus*, which are widely distributed in tropical and sub-tropical regions [4]. In the past few years, several studies have reported concurrent outbreaks of DENV, CHIKV and ZIKV in the same geographical area [5, 6]. Currently, unprecedented outbreaks of DENV, CHIKV and ZIKV are co-occurring in Brazil. In 2017, the Brazilian Ministry of Health estimated that approximately 251,000 suspected cases of DENV, 185,000 suspected cases of CHIKV and close to 18,000 suspected ZIKV cases had occurred in Brazil [7].

Monitoring virus genotype diversity is crucial to understand the emergence and spread of outbreaks, both aspects that are vital to develop effective prevention and treatment strategies. Both DENV and CHIKV epidemics are associated with a mortality and morbidity that puts a significant economic burden on the affected regions [8,9]. While infections with ZIKV are rarely fatal, as stated before, ZIKV infections may result in Guillain-Barré syndrome and congenital malformations [10,11]. Genomic surveillance of epidemics at the appropriate resolution and consistently classifying the reported genetic sequences, also enables the identification of strains associated with greater epidemic potential [12] or disease severity [13].

However, methods that consistently classify arbovirus sequences at the level of species and sub-species (i.e. serotype and/or genotype) are currently lacking. Additionally, whole genome sequences are often not available in routine clinical settings, forcing the use of shorter gene sequences to classify at viral species or sub-species level. It has however insufficiently been explored which genomic regions are most suitable for accurate classification.

A new computational method for the identification of DENV/CHIKV/ZIKV sequences, with respect to species and sub-species (i.e. serotype and/or genotype), is presented. The classification method is implemented in the Genome Detective software tool, which was validated on a large dataset by assessing the classification performance of whole-genome sequences, partial-genome sequences and products from next-generation sequencing methods. Furthermore, the suitability of different genomic regions for virus classification was evaluated.

Materials and methods

Datasets

Global whole-genome sequence dataset (Global-WG). A dataset of previously published whole-genome sequences from GenBank [14] was compiled. This dataset consists out of 4,118 DENV sequences, 653 CHIKV sequences and 413 ZIKV sequences and contains DENV

sequences for each of the four known serotypes: DENV-sero1 (n = 1688), DENV-sero2 (n = 1317), DENV-sero3 (n = 897) and DENV-sero4 (n = 216). The list of GenBank accession numbers for this global whole-genome dataset is available in the Supporting Information section ([S1 File](#)). In the remainder of this manuscript, this dataset will be referred to as **Global-WG**.

Global envelope sequence dataset (Global-ENV). A dataset of previously published envelope sequences from GenBank [14] was compiled. This dataset consists out of 4,118 DENV sequences, 2,531 CHIKV sequences and 413 ZIKV sequences and contains DENV sequences for each of the four known serotypes: DENV-sero1 (n = 1688), DENV-sero2 (n = 1317), DENV-sero3 (n = 897) and DENV-sero4 (n = 216). The list of GenBank accession numbers for this global envelope dataset is available in the Supporting Information section ([S2 File](#)). In the remainder of this manuscript, this dataset will be referred to as **Global-ENV**.

Identification of genotypes and selection of reference sequences. To identify the viral genotypes, a multiple sequence alignment was constructed with the MAFFT alignment software [15] per virus species, using the **Global-WG** dataset. Each alignment was edited manually until a codon-correct alignment was achieved in all genes. The next step in this exploration involved a phylogenetic analysis using PhyML (i.e. Maximum likelihood, 1000 bootstrap replicates) and MrBayes (i.e. Bayesian) [16,17]. With this approach, four main DENV clades (i.e. serotype 1 to 4) and 19 genotypes (i.e. 1I, 1II, 1III, 1IV, 1V, 2I, 2II, 2III, 2IV, 2V, 2VI, 3I, 3II, 3III, 3V, 4I, 4II, 4III and 4IV) were identified. These findings are in agreement with the current consensus in DENV classification [18–21]. For CHIKV, three phylogenetic clades can be distinguished: The East-Central-South African (ECSA) genotype, the Asian-Caribbean genotype and the West African genotype. The West African genotype being more divergent and less widespread than the ECSA genotype and the Asian-Caribbean genotype [22,23]. ZIKV, as well, can be classified into two genotypes. The African genotype, originally identified in Uganda in 1947 [24], is found in many African countries [25]. The Asian genotype was identified in Malaysia in 1966 [26], this genotype has recently caused the worldwide epidemic in Asia and the Pacific [27,28], and is responsible for the epidemic in the Americas [5].

The accuracy and consistency with which a method identifies viral species and genotype clades depends on the selection of a set of representative reference sequences [29–31].

The initial step in the selection of reference strains for our method involved the identification of highly divergent but equidistant whole-genome sequences that are representative for the diversity within the different DENV, CHIKV and ZIKV genotypes, by screening all published complete genome sequences in our **Global-WG** dataset. For example, we normally start by selecting 5–10 sequences that represent the diversity of each virus genotypes. Sequences that met these selection criteria were quality controlled for the presence of insertions, deletions, frame shifts and non-IUPAC characters using VIRULIGN [32]. For DENV, we used the reference sequences that are included with the VIRULIGN software, for ZIKV, we used the reference sequence presented in [33], and for CHIKV we constructed a new reference sequence from NC_004162 that we added to the VIRULIGN repository. Sequences that pass the quality control were aligned using MAFFT [15], and were subjected to phylogenetic analysis using PAUP* (i.e. Neighbor Joining), MrBayes (i.e. Bayesian) and PhyML (i.e. Maximum likelihood) [16,17,34,35] using GTR+G+I. Sequences that gave consistent topologies using all three tree inference methods were retained as potential reference sequences (see [Supporting Information, S1 Table](#)) and used in the next step of the evaluation process.

We established that none of the selected reference strains were recombinants ([S2 Fig](#)) using the recombination detection program RDP4 [36]

Suitability of sub-genomic regions for genotyping purposes. The reference strain dataset ([S1 Table](#)) was then explored to establish the suitability of sub-genomic regions for automated genotyping. Two different methods were used.

The first was a boot-scanning method, using a sliding window approach exploring the range between 200 and 2,000 nucleotides. All windows across the genome were used for the construction of Neighbor joining trees with 1,000 bootstrap replicates. The aim was to find the size and segments of the genome that would correctly classify a query sequence with a bootstrap support of >70%.

The second method involved the calculation of the phylogenetic signal present in each of the DENV, CHIKV and ZIKV genes, using the same set of reference sequences. To compute the phylogenetic signal, the TreePuzzle software [37] implementation of the likelihood-mapping method [38] was used. Only between-genotype quartets were evaluated. Quartet puzzling essentially is a three-step procedure, first reconstructing all possible quartet maximum likelihood trees (maximum-likelihood step), then repeatedly combining the quartet trees to an overall tree (puzzling step), and finally computing the majority rule consensus of all intermediate trees giving the quartet puzzling tree (consensus step).

Classification method and implementation

Classification method. Our method involves a viral classification pipeline, drawing inspiration from the one described previously to classify HIV, hepatitis C virus and human T-lymphotropic virus sequences [29,30]. The classification pipeline presented here consists of two classification components. The first classification component enables species and sub-species assignments. The classification analysis subjects a query sequence to a BLAST analysis against a set of reference sequences [39]. A query is assigned to a particular type when BLAST reports an assignment with a score that exceeds a predefined threshold.

The second classification component involves the construction of a Neighbor Joining phylogenetic tree. This component enables assignments on genotype and/or subtype level. First, the query sequence is aligned with a set of reference sequences.

The alignment is produced using the profile alignment option in the ClustalW software [40], such that the query sequence is added to the existing alignment of reference sequences. Subsequent to the alignment, a Neighbor Joining phylogenetic tree, with 100 bootstrap replicates, is constructed. The tree is constructed using the HKY distance metric with gamma among-site rate variation, as implemented in the PAUP* software [34]. The query sequence is assigned to a particular genotype if it clusters monophyletically with that genotype clade with bootstrap support >70%. If the bootstrap support is <70%, the genotype is reported to be unassigned.

Software implementation. While the classification method was inspired by the one previously presented [29], a new software framework was developed to be easily adaptable to the classification procedures for various viral pathogens. All source code is written in the Java programming language (Fig 1). The software framework is part of the Genome Detective tool-chain [41]

ArboTyping classification method and implementation. Firstly, the viral species is determined using BLAST, classifying the sequence as DENV, CHIKV or ZIKV.

In case the submitted sequence was assigned either as ZIKV or CHIKV, a Neighbor joining tree is inferred to determine the respective ZIKV or CHIKV genotype. Only for DENV, another BLAST procedure is invoked to assign the serotype first. Based on the inferred serotype, a serotype specific Neighbor joining tree is constructed to determine the Dengue genotype.

For each of these steps, the earlier discussed reference strains were used, with respect to the appropriate typing level (i.e. virus species, serotype or genotype). This process is summarized in a decision tree in Fig 2.

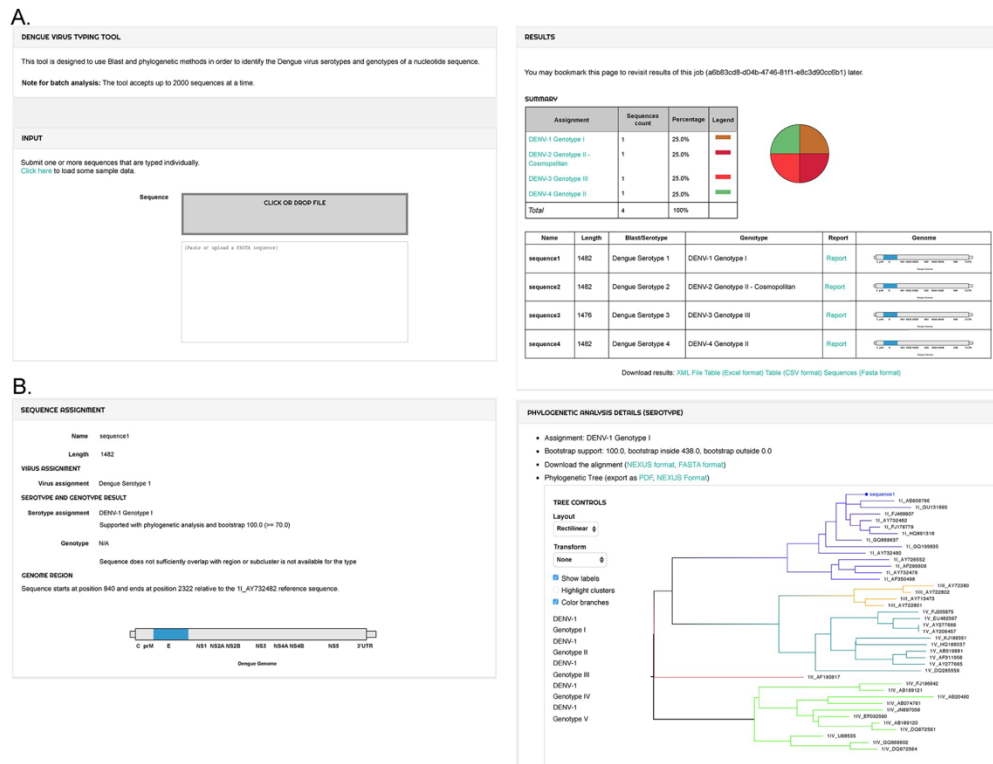


Fig 1. The typing tools' web interface. The web interface provides users a portal to run classifications on their sequences and to visualize the classification results (A). The typing report presents information about the sequence name of the query sequence, the nucleotide length of the sequence, an illustration of the position of the sequence in the virus' genome, the species assignment and the genotype assignment. A detailed report is provided for the phylogenetic analysis that resulted into this classification. All results can be exported to a variety of file formats (XML, CSV, Excel or FASTA format). The detailed HTML report (B) contains information on the sequence name, length, assigned virus and genotype, an illustration of the position of the sequence in the virus' genome and the phylogenetic analysis section. The phylogenetic analysis section shows the alignment and constructed phylogeny: the query sequence is always shown at the top of the phylogenetic tree.

<https://doi.org/10.1371/journal.pntd.0007231.g001>

Testing revealed that a BLAST cut-off value of 200 allowed accurate identification of the virus species and DENV serotypes using sequence segments >150 base pairs.

Note that the species and serotype classification procedure are implemented as separate BLAST steps. This enables the tool to efficiently perform large throughput species classification, such as for the classification of next-generation sequencing reads.

An instance of the ArboTyping web application is publically available on a dedicated server (<http://krisp.org.za/tools.php>). The web interface on this server accepts up to 2,000 whole-genome or partial genome sequences at a time. The tool can be accessed by the Genome Detective interface or by the selection of individual viruses typing tool (i.e. Zika, Dengue and Chikungunya).

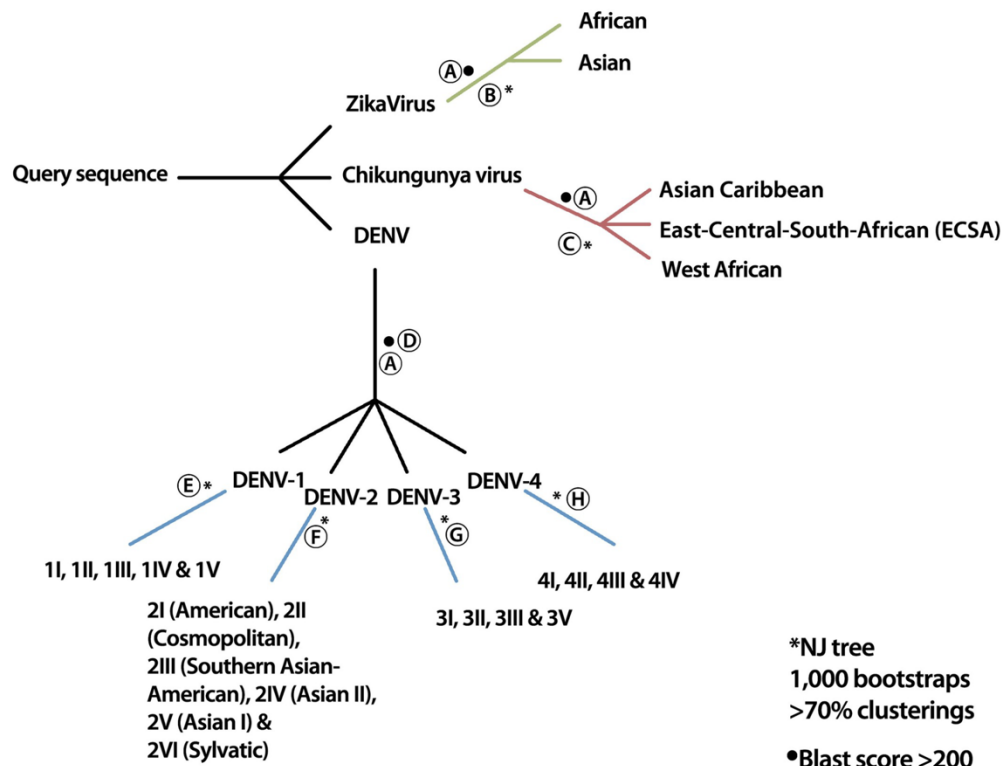


Fig 2. Outline of the classification procedure. Firstly (A), the viral species is determined using BLAST. When the submitted sequence is a *Zika virus*, a Neighbor joining tree is constructed to determine the Zika genotype (B). When the submitted sequence is a *Chikungunya virus*, a Neighbor joining tree is constructed to determine the Chikungunya genotype (C). When the submitted sequence is a *Dengue virus*, the serotype is determined using another BLAST invocation (D). Based on the inferred serotype, a serotype specific Neighbor joining tree is constructed to determine the Dengue genotype (E, F, G, H).

<https://doi.org/10.1371/journal.pntd.0007231.g002>

Classification performance for whole-genomes and sub-genomic regions. To determine the accuracy of the automated method for whole-genome sequences, the method was evaluated on a whole-genome sequence dataset (i.e. **Global-WG** dataset).

As sequences from sub-genomic regions are more commonly available than whole-genome sequences, the method's accuracy was also evaluated in this context. For this purpose, the envelope sequences in the **Global-ENV** dataset were used for evaluation.

Each of the sequences considered for evaluation was assigned using both the gold standard and the here described automated method. The gold standard, a manual classification consists of performing an assignment using both Bayesian (i.e. MrBayes, assignment with posterior > 90% [17]) and Maximum likelihood (i.e. PhyML, 1000 bootstrap replicates, assignment with > 70% of replicates [16]) phylogenetic analysis. When the assignments generated by both the Bayesian and Maximum likelihood technique match, the classification is confirmed [31].

The sensitivity, specificity and accuracy of our method was calculated for both species assignment and genotyping. Sensitivity was computed by the formula $\frac{TP}{TP+FN}$, specificity by the formula $\frac{TN}{TN+FP}$ and accuracy by the formula $\frac{TP+TN}{TP+FP+FN+TN}$ [42]. In these formulas: TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives.

Results

ArboTyping classification method and implementation

An efficient method to classify virus sequences with respect to their species and sub-species (i.e. serotype and/or genotype) was developed. This method was implemented in Java and this implementation was integrated in an easy-to-use web interface. A detailed description of the method and its implementation can be found in the 'Classification method and implementation' Methods subsection.

Suitability of sub-genomic regions for genotyping purposes

Two different methods were used to verify the suitability of sub-genomic regions for genotyping purposes: a boot-scanning method and a likelihood-mapping method (see [Methods](#)).

For DENV, the only sub genomic region that supports confident genotype assignment across the four different serotypes was the envelope gene. For CHIKV, the envelope region E1 was the only region that allowed consistent assignment. The boot-scanning analysis showed that for ZIKV, segments of around 1,200–1,500 base pairs support the genotype assignment with bootstrap > 70% ([Fig 3](#)). This was the case over the entire genome, with the exception of the end of the genome (i.e. the non-coding region) and near the NS3 region, where bootstraps fell below 60%.

Our likelihood-mapping analyses show that for DENV, the envelope, NS1, NS3 and NS5 had good phylogenetic signal across all four serotypes. For CHIKV, the envelope E2 gene had the best signal but this region did not provide good boot-scanning support for the classification of the ECSA genotype ([Fig 3](#)). For ZIKV, the envelope, NS1, NS2A, NS3, NS4A, NS4B and NS5 regions had good phylogenetic signal. A detailed overview of the results of the likelihood-mapping analysis can be found in the S2 Table of the Supporting Information.

In summary, these analyses show that the envelope genes of the reference datasets of the three pathogens (DENV, 1,485 nucleotides; CHIKV, 1,317 nucleotides; ZIKV, 1,525 nucleotides) are the most suitable targets for reliable genotype classification.

Classification performance for whole-genome sequences

Our automated method provided specificity, sensitivity and accuracy of 100% for the identification of complete genomes for all viral species and genotypes compared to the gold standard, a manual classification. For a detailed overview of the DENV, CHIKV and ZIKV assignment performance, we refer to the Supporting Information [S3 Table](#).

Only ten of 4118 DENV whole-genomes could not be classified at the genotype level, either by manual phylogenetic analysis or by our automated method. Notably, the seven sequences (AF298807, KF864667, EU179860, JQ922546, KF184975, KF289073, EF457905 of DENV-Sero1 were outliers in the phylogenetic tree (see [Supporting Information, S1 Fig](#)). We tested all ten sequences for recombination using boot-scanning (see [Supporting Information, S2 Fig](#)) and the recombination detection program RDP4 [36]. We only found sequence AY496879 to be a clear recombinant of DENV genotype 3I and 3II. The two other sequences (DENV-Sero2 KF744408 and DENV-Sero3 JF262783) were also identified as a divergent outlier.

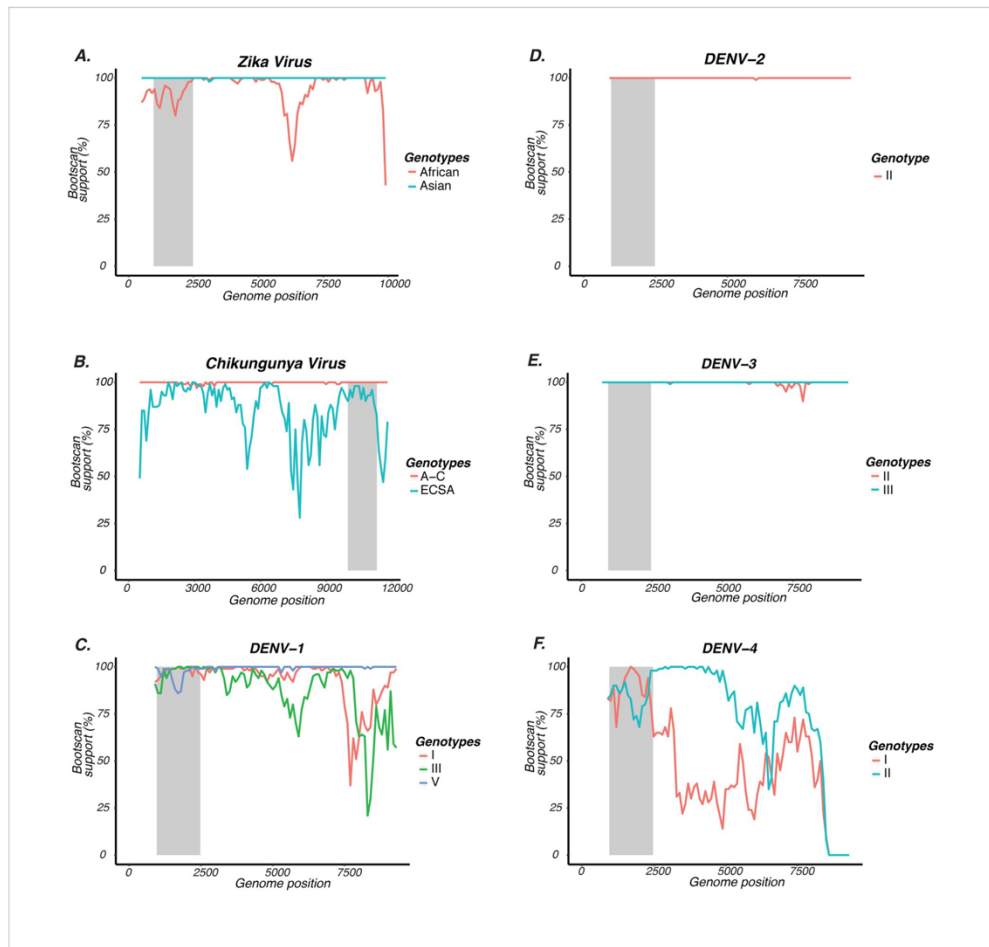


Fig 3. Investigating the suitability of sub-genomic regions for genotyping through boot-scanning. This plot was constructed using bootstrap results from Neighbor Joining trees (1000 bootstrap replicates), performed on the dataset of the indicated reference strains. The boot-scanning method uses a sliding window of a 1500 base pair segment that moves with steps of 100 base pairs along the genome. The X-axis represents the nucleotide position in the genome, and the Y-axis represents bootstrap support in percentages. The light-grey rectangular area marks the location of the envelope gene in each graph. The inset color code shows the genotypes defined in each virus species. For ZIKV (A), this is African and Asian. For CHIKV (B), this is Asian Caribbean (A-C), West African (WA) and East-Central-South African (ECSA). For DENV (C-F), the genotype is visualized by a Roman numeral. Only genotypes which showed less than 100% bootstrap support across the genome are shown.

<https://doi.org/10.1371/journal.pntd.0007231.g003>

Classification performance for sub-genomic regions

Our analysis shows that the classification results for the envelope sub-genomic region at the species and genotype level were similar to that obtained using whole-genome sequences and largely in agreement with the gold standard, a manual classification.

For DENV, most of the genotypes were classified with great accuracy (i.e. specificity and sensitivity > 99%) using the envelope gene. The exception was DENV-sero2 genotype IV, of which 41 envelope sequences were available and for which 33 were correctly identified (i.e. sensitivity 80.49%, specificity 100%). The CHIKV sequences covering the E1 region were accurately classified for all three genotypes (i.e. 100% sensitivity and specificity). All the ZIKV envelope sequences were classified with 100% sensitivity and specificity. For a detailed overview of the DENV, CHIKV and ZIKV assignment performance refer to Supporting Information [S4 Table](#).

Since a good phylogenetic signal was reported for the DENV and ZIKV NS5 region and the CHIKV E2 region, a classification analysis was performed for these regions as well. For the DENV NS5 region a sensitivity of 57.48% and specificity of 31.35% was observed. Nearly all ZIKV NS5 sequences were correctly assigned to the African genotype (i.e. sensitivity of 97.72% and specificity of 100%). This indicates that the ZIKV NS5 region might also be used for genotype classification. For CHIKV, the E2 region showed perfect accuracy, similar to the E1 region (i.e. specificity and sensitivity of 100%). However, our previous boot-scanning support showed that the genetically variable E2 region may cause problems for some strains to be correctly identified as ECSA genotype.

In summary, our results suggest that the envelope region of DENV and ZIKV and the E1 envelope region of CHIKV are suitable for genotyping purposes. In addition, these regions contain the largest number of sequences in public databases, which easily allows for a wide range of comparative analyses and validation experiments.

Discussion

Emerging infectious diseases caused by viral pathogens still represent a major threat to public health worldwide, as recently demonstrated by outbreaks of Ebola, Zika, Middle East Respiratory Syndrome (MERS) and Yellow Fever virus. Fast and accurate real-time monitoring of outbreaks and surveillance of on-going epidemics is crucial to anticipate viral spread and to design effective prevention or treatment strategies. To this end, an accurate and reliable method for the classification of ZIKV/DENV/CHIKV arboviruses was developed: The ArboTyping tool.

The ArboTyping tool implements a classification pipeline that consists of a BLAST-based species assignment and phylogenetic assessment to identify subspecies (i.e. genotypes) with respect to a set of reference strains, as exemplified for other virus species by previous work [29–31]. To enable accurate classification, a set of reference sequences that cover the extent of diversity within species and subspecies, was carefully selected.

The classification performance of the ArboTyping tool was assessed on a dataset of whole-genome sequences. All whole-genome sequences in this dataset that could be confidently assigned a species and genotype with the gold standard, a manual classification procedure, were concordant with the typing tool.

There were, however, 10 sequences that could not be classified using the manual classification procedure: further analyses show that these 10 sequences consist out of 3 outlier sequences, 2 clades of outlier sequences (3 sequences in each outlier clade) and 1 recombinant sequence. As these outliers have been previously identified [43], these results need to be further investigated to assess whether these outliers form new genotypes [44].

However, whole-genome sequences are currently not routinely available and the suitability of the different genomic regions was evaluated with respect to their use for classification. Since the envelope gene is a popular target for phylogenetic classification, there is a large availability of envelope sequences in public databases. Therefore, the performance of the ArboTyping tool was evaluated on a large dataset of envelope sequences (i.e. **Global-ENV** dataset). For these envelope sequences, a classification performance close to the tool's performance on whole-genome sequences was reported.

While the availability of sequence products originating from other genomic regions is currently low, it can be expected that these regions will increase in relevance given the interest in developing antiviral agents that target non-structural proteins. Therefore, more detailed studies to assess the classification performance of other genomic regions are warranted [44].

In this manuscript, we focus on the classification of consensus sequences on the species and sub-species level. However, Genome Detective, the framework in which our tools are integrated, is also a virus discovery toolchain [41]. Genome Detective's user interface allows users to supply raw next-generation sequence reads that can be automatically assembled into a consensus and passed to the ArboTyping tool. Details on the methods used to assemble reads in Genome Detective and an extensive validation using raw NGS reads can be found in [41].

In conclusion, the new method presented here allows the fast, accurate and high-throughput classification of DENV, CHIKV and ZIKV species and genotypes. Species can be classified using different sequencing products (i.e. whole-genome sequences, envelope sequences and individual next-generation sequencing reads) and genotypes can be classified most confidently when using envelope sequences or whole-genome sequences. This method accommodates the need to consistently and accurately classify DENV/CHIKV/ZIKV sequences, which is essential to implement epidemic tracing and to support outbreak surveillance efforts. Additionally, we present a solid framework that has the potential to serve as the foundation for many other arbovirus classification tools. These tools are also useful to be integrated in data management environments [45].

Our method is implemented in the Genome Detective software framework, suitable for many virus typing tools. The web application that makes our tool available through an easy-to-use web interface is available online via a dedicated server that is hosted at <http://www.krisp.org.za/tools.php>.

Supporting information

S1 Fig. Maximum likelihood phylogenetic tree of the DENV-sero1 outliers. All full genome DENV-sero1 sequences were assigned to genotype-level using manual phylogenetic analysis and classification by the automated typing tool. In total, seven full genomes of DENV-sero1 could not be classified at genotype level by either classification method. These seven sequences are visualized in a phylogenetic tree of the WGS datasets, colored according to genotype. (II in blue, III in green, IIII in red, IIIV in yellow, IV in pink) It can be seen that a divergent cluster of six genomes (AF298807, KF864667, KF184975, EU179860, KF289073 and JQ922546 in black) form an outlier clade and one genome (EF457905 in black) can be considered an outlier. However, note that these seven genomes could be properly assigned to serotype 1. (TIF)

S2 Fig. Recombination analysis for the DENV whole genome sequences. The bootscan results for the ten whole genomes of DENV that could not be classified at genotype level are shown. Boot-scanning analysis was performed using a window length of 1500 base pairs and a step size of 100 base pairs. The different colours represent the genotypes for each serotype. The X-axis represents the nucleotide position in the genome and the Y-axis represents bootstrap results in

percentages. In total, 7 DENV-sero1 sequences were analysed and 1 sequence for each of the other serotypes, i.e. DENV-sero2, DENV-sero3 and DENV-sero4. We only found sequence AY496879 to be a recombinant of DENV genotype 3I and 3II. The other sequences are outliers (i.e. JF262783, KF744408, EF457905) or clades of outliers (i.e.: AF298807, KF864667 and KF184975 form an outlier clade; EU179860, KF289073 and JQ922546 form an outlier clade). (TIF)

S1 Table. Reference strains selected for the DENV, CHIKV, ZIKV genotypes. These reference sequences were selected to be representative for the diversity within the different DENV, CHIKV and ZIKV genotypes that circulate within these virus species. (DOCX)

S2 Table. Phylogenetic signal estimated by likelihood mapping for DENV (DENV-sero1 to DENV-sero4), CHIKV and ZIKV sub-genomic regions. Phylogenetic signal was calculated separately per protein by the likelihood mapping method implemented in the software Tree-Puzzle. Likelihood mapping analysis computes the likelihood of the three possible trees that can be constructed from all possible inter-genotype quartets of taxa. The results for the resolved quartets and unresolved quartets are shown in the table, while the partially resolved quartets are not listed (can be obtained by 100%—(un)resolved quartets). Partially resolved quartets represent the quartets for which conflicting phylogenetic signal or potential recombination is present. Genomic regions for which the percentage of resolved quartets is higher than 90% are shaded in orange and are considered to be characterized by sufficient phylogenetic signal. (DOCX)

S3 Table. Evaluation of the automated phylogenetic method to classify DENV, CHIKV and ZIKV whole-genome genomes. The new classification method consists of 2 parts: determining the species (and for DENV also the serotype) using a BLAST procedure, followed by determining the genotype using an automated phylogenetic method. Our method was able to assign all sequences in the whole-genome validation dataset to the right species and DENV serotype. Therefore, in this table, we focus on the classification performance with respect to genotype assignment, based on the output of the BLAST step (i.e. a dataset of the proper species and serotype). The classification results were compared to manual phylogenetic analysis. Column names: TP = total positives, TN = total negatives, FP = false positive, FN = false negative, SENS = sensitivity, SPEC = specificity, ACC = accuracy. (DOCX)

S4 Table. Evaluation of the automated phylogenetic method to classify DENV, CHIKV and ZIKV envelope genomes. The new classification method consists of 2 parts: determining the species (and for DENV also the serotype) using a BLAST procedure, followed by determining the genotype using an automated phylogenetic method. Our method was able to assign all sequences in the envelope validation dataset to the right species and DENV serotype. Therefore, in this table, we focus on the classification performance with respect to genotype assignment, based on the output of the BLAST step (i.e. a dataset of the proper species and serotype). The classification results were compared to manual phylogenetic analysis. Column names: TP = total positives, TN = total negatives, FP = false positive, FN = false negative, SENS = sensitivity, SPEC = specificity, ACC = accuracy. (DOCX)

S1 File. Accession number of the sequences collected from DENV, ZIKV and CHIKV whole-genome genomes. A GenBank mining of sequences was performed against whole-

genome genomes of these viruses that had the genotype reported for sensitivity, specificity and accuracy tests of the tool.

(XLSX)

S2 File. Accession number of the sequences collected from DENV, ZIKV and CHIKV envelope genomes. A GenBank mining of sequences was performed against envelope genomes of these viruses that had the genotype reported for sensitivity, specificity and accuracy tests of the tool.

(XLSX)

Author Contributions

Conceptualization: Vagner Fonseca.

Data curation: Vagner Fonseca, Pieter J. K. Libin, Lize Cuypers.

Formal analysis: Vagner Fonseca, Pieter J. K. Libin, Kristof Theys, Maria I. Restovic, Murilo Freire, Lize Cuypers.

Funding acquisition: Luiz C. J. Alcantara, Tulio de Oliveira.

Investigation: Vagner Fonseca, Pieter J. K. Libin, Lize Cuypers.

Methodology: Vagner Fonseca, Maria I. Restovic, Murilo Freire.

Project administration: Luiz C. J. Alcantara, Tulio de Oliveira.

Resources: Luiz C. J. Alcantara, Tulio de Oliveira.

Software: Vagner Fonseca, Kristof Theys, Maria I. Restovic, Murilo Freire, Koen Deforche, Emmanuel J. San, Kaliane C. B. Machado.

Supervision: Luiz C. J. Alcantara, Tulio de Oliveira.

Validation: Vagner Fonseca, Pieter J. K. Libin, Kristof Theys, Maria I. Restovic, Murilo Freire, Lize Cuypers, Koen Deforche, Anne-Mieke Vandamme.

Writing – original draft: Vagner Fonseca, Nuno R. Faria, Marcio R. T. Nunes, Marta Giovanetti, Ann Nowé, Ana Abecasis, Koen Deforche, Gilberto A. Santiago, Isadora C. de Siqueira, Vasco Azevedo, Ana Maria Bispo-de Filippis, Rivaldo Venâncio da Cunha, Oliver G. Pybus, Anne-Mieke Vandamme, Luiz C. J. Alcantara, Tulio de Oliveira.

Writing – review & editing: Vagner Fonseca, Nuno R. Faria, Marcio R. T. Nunes, Marta Giovanetti, Ann Nowé, Ana Abecasis, Koen Deforche, Gilberto A. Santiago, Isadora C. de Siqueira, Emmanuel J. San, Vasco Azevedo, Ana Maria Bispo-de Filippis, Rivaldo Venâncio da Cunha, Oliver G. Pybus, Anne-Mieke Vandamme, Luiz C. J. Alcantara, Tulio de Oliveira.

References

1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. *Nature*. 2013; 496:504–507. <https://doi.org/10.1038/nature12060> PMID: 23563266
2. Weaver SC, Lecuit M. Chikungunya Virus and the Global Spread of a Mosquito-Borne Disease. *New England Journal of Medicine*. 2015; 372(13):1231–1239. <https://doi.org/10.1056/NEJMra1406035> PMID: 25806915
3. Fauci AS, Morens DM. Zika Virus in the Americas—Yet Another Arbovirus Threat. *New England Journal of Medicine*. 2016; 374(7):601–604. <https://doi.org/10.1056/NEJMp1600297> PMID: 26761185
4. Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. Albopictus*. *eLife*. 2015; 4(08347). <https://doi.org/10.7554/eLife.08347> PMID: 26126267

5. Cardoso CW, Paploski IAD, Kikuti M, Rodrigues MS, Silva MMO, Campos GS, et al. Outbreak of Exanthematous Illness associated with Zika, Chikungunya, and Dengue viruses, Salvador, Brazil. *Emerging Infectious Diseases*. 2015; 21(12):2274–2276. <https://doi.org/10.3201/eid2112.1511167> PMID: 26584464
6. Roth A, Mercier A, Lepers C, Hoy D, Duituturaga S, Benyon E, 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*. 2014; 19(41):20929. <http://dx.doi.org/10.2807/1560-7917.ES2014.19.41.20929>. PMID: 25345518
7. Ministério de Saúde B. Boletim Epidemiológico Secretaria de Vigilância em saúde; 2018. v. 49.
8. Shepard DS, Undurraga EA, Halasa YA, Stanaway JD. The global economic burden of dengue: A systematic analysis. *Lancet Infect Dis*. 2016; 16(8):935–941. [https://doi.org/10.1016/S1473-3099\(16\)00146-6](https://doi.org/10.1016/S1473-3099(16)00146-6) PMID: 27091092
9. Morens DM, Fauci AS. Meeting the Challenge of Epidemic Chikungunya. *Journal of Infectious Diseases*. 2016; 214(suppl 5):S434–S435. <https://doi.org/10.1093/infdis/jiw291> PMID: 27920168
10. Rasmussen SA, Jamieson DJ, Honein MA, Petersen LR. Zika Virus and Birth Defects—Reviewing the Evidence for Causality. *New England Journal of Medicine*. 2016; p. 1–7. <https://doi.org/10.1056/NEJMs1604338> PMID: 27074377
11. Brasil P, Sequeira PC, Freitas AD, Zogbi HE, Calvet GA, de Souza RV, et al. Guillain-Barré syndrome associated with Zika virus infection. *The Lancet*. 2016; 1482. [https://doi.org/10.1016/S0140-6736\(16\)30058-7](https://doi.org/10.1016/S0140-6736(16)30058-7).
12. Manokaran G, Finol E, Wang C, Gunaratne J, Bahl J, Ong EZ, et al. Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. *Science*. 2015; 350(6257):217–221. <https://doi.org/10.1126/science.aab3369> PMID: 26138103
13. Katzelnick LC, Fonville JM, Gromowski GD, Bustos Arriaga J, Green A, James SL, et al. Dengue viruses cluster antigenically but not as discrete serotypes. *Science (New York, NY)*. 2015; 349(6254):1338–43. <https://doi.org/10.1126/science.aac5017> PMID: 26383952
14. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. *Nucleic acids research*. 2013; 41(D1):D36–D42. <https://doi.org/10.1093/nar/gkr1202>
15. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. 2013; 30(4):772–780. <https://doi.org/10.1093/molbev/mst010> PMID: 23329690
16. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*. 2003; 52(5):696–704. <https://doi.org/10.1080/10635150390235520> PMID: 14530136
17. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003; 19(12):1572–1574. <https://doi.org/10.1093/bioinformatics/btg180> PMID: 12912839
18. Rico-Hesse R. Molecular evolution and distribution of dengue viruses type 1 and 2 in nature. *Virology*. 1990; 174(2):479–493. [https://doi.org/10.1016/0042-6822\(90\)90102-W](https://doi.org/10.1016/0042-6822(90)90102-W) PMID: 2129562
19. Twiddy SS, Farrar JJ, Vinh Chau N, Wills B, Gould EA, Gritsun T, et al. Phylogenetic relationships and differential selection pressures among genotypes of dengue-2 virus. *Virology*. 2002; 298(1):63–72. <https://doi.org/10.1006/viro.2002.1447> PMID: 12093174
20. Chungue E, Deubel V, Cassar O, Laille M, Martin PMV. Molecular epidemiology of dengue 3 viruses and genetic relatedness among dengue 3 strains isolated from patients with mild or severe form of dengue fever in French Polynesia. *Journal of general virology*. 1993; 74(12):2765–2770. <https://doi.org/10.1099/0022-1317-74-12-2765>
21. Klungthong C, Zhang C, Mammen MP, Ubol S, Holmes EC. The molecular epidemiology of dengue virus serotype 4 in Bangkok, Thailand. *Virology*. 2004; 329(1):168–179. <https://doi.org/10.1016/j.virol.2004.08.003> PMID: 15476884
22. Nunes MRT, Faria NR, de Vasconcelos JM, Golding N, Kraemer MU, de Oliveira LF, et al. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Medicine*. 2015; 13(1):102. <https://doi.org/10.1186/s12916-015-0348-x> PMID: 25976325
23. Volk SM, Chen R, Tsetsarkin KA, Adams AP, Garcia TI, Sall AA, et al. Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. *Journal of virology*. 2010; 84(13):6497–6504. <https://doi.org/10.1128/JVI.01603-09> PMID: 20410280
24. Dick GWA, Kitchen SF, Haddock AJ. Zika virus (I). Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 1952; 46(5):509–520. [https://doi.org/10.1016/0035-9203\(52\)90042-4](https://doi.org/10.1016/0035-9203(52)90042-4) PMID: 12995440

25. Kindhauser MK, Allen T, Frank V, Santhana RS, Dye C. Zika: the origin and spread of a mosquito-borne virus. *Bull World Health Organ*. 2016;171082.
26. Marchette NJ, Garcia R, Rudnick A. Isolation of Zika virus from *Aedes aegypti* mosquitoes in Malaysia. *American Journal of Tropical Medicine and Hygiene*. 1969; 18(3):411–415. <https://doi.org/10.1056/NEJMp1002530> PMID: 4976739
27. Cao-Lormeau VM, Roche C, Teissier A, Robin E, Berry AL, Mallet HP, et al. Zika Virus, French Polynesia, South Pacific, 2013. *Emerging Infectious Diseases*. 2014; 20(6):1085–1086. <https://doi.org/10.3201/eid2006.140138> PMID: 24856001
28. Hayes EB, Others. Zika virus outside Africa. *Emerg Infect Dis*. 2009; 15(9):1347–1350. <https://doi.org/10.3201/eid1509.090442> PMID: 19788800
29. Alcantara LCJ, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Research*. 2009; 37(Suppl 2):634–642. <https://doi.org/10.1093/nar/gkp455> PMID: 19483099
30. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*. 2005; 21(19):3797–3800. <https://doi.org/10.1093/bioinformatics/bti607> PMID: 16076886
31. Pineda-Peña AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infection, Genetics and Evolution*. 2013; 19:337–348. <https://doi.org/10.1016/j.meegid.2013.04.032> PMID: 23660484
32. Libin P., Deforche K., Abecasis A. B., & Theys K. (2018). VIRULIGN: fast codon-correct alignment and annotation of viral genomes. *Bioinformatics (Oxford, England)*.
33. Theys K., Libin P., Dallmeier K., Pineda-Peña A. C., Vandamme A. M., Cuypers L., & Abecasis A. B. (2017). Zika genomics urgently need standardized and curated reference sequences. *PLoS pathogens*, 13(9), e1006528. <https://doi.org/10.1371/journal.ppat.1006528> PMID: 28880955
34. Salemi M, Vandamme AM. *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press; 2003.
35. Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*. 2008; 24(4):581–583. <https://doi.org/10.1093/bioinformatics/btm388> PMID: 17766271
36. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuve P. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics*. 2010; 26(19):2462–2463. <https://doi.org/10.1093/bioinformatics/btq467> PMID: 20798170
37. Ha Schmidt, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics (Oxford, England)*. 2002; 18(3):502–504. <https://doi.org/10.1093/bioinformatics/18.3.502>
38. Strimmer K, von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences*. 1997; 94(13):6815–6819. <https://doi.org/10.1073/pnas.94.13.6815> PMID: 9192648
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool; 1990.
40. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007; 23(21):2947–2948. <https://doi.org/10.1093/bioinformatics/btm404> PMID: 17846036
41. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, Alcantara LC, Vanden Eynden E, Vandamme AM, Deforche K, de Oliveira T, *Bioinformatics* (2019), <https://doi.org/10.1093/bioinformatics/bty695> PMID: 30124794
42. Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, et al. Evaluation of diagnostic tests for infectious diseases: General principles. *Nature Reviews Microbiology*. 2006; 4(9 SUPPL.) S21–S31. <https://doi.org/10.1038/nrmicro1523> PMID: 17034069
43. Libin P., Vanden Eynden E., Incardona F., Nowé A., Bezenchek A., EucoHIV Study Group, Sönnnerborg A., Vandamme A.-M., Theys K., Baele G.(2017). PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context. *Bioinformatics*, 33(24):3993–3995. <https://doi.org/10.1093/bioinformatics/btx535> PMID: 28961923
44. Cuypers L., Libin P.J.K., Simmonds P., Nowé A., Muñoz-Jordán, J., Alcantara L.C.J., Vandamme A.-M., Santiago G.A., Theys K. (2018). Time to Harmonize Dengue Nomenclature and Classification. *Viruses*, 10(10), pii: E569. <https://doi.org/10.3390/v10100569> PMID: 30340326
45. Libin P., Beheydt G., Deforche K., Imbrechts S., Ferreira F., Van Laethem K., Theys K., Carvalho A.P., Cavaco-Silva J., Lapadula G., Torti C., Assel M., Wesner S., Snoeck J., Ruelle J., De Bel A., Lacor P.,

De Munter P., Van Wijngaerden E., Zazzi M., Kaiser R., Ayoub A., Peeters M., de Oliveira T., Alcantara L.C., Grossman Z., Sloom P., Otelea D., Paraschiv S., Boucher C., Camacho R.J., Vandamme A.-M. (2013). RegaDB: community-driven data management and analysis for infectious diseases. *Bioinformatics*. 2013, 29(11):1477–80. <https://doi.org/10.1093/bioinformatics/btt162> PMID: 23645815

CHAPTER 4: GENOME DETECTIVE CORONAVIRUS TYPING TOOL FOR RAPID IDENTIFICATION AND CHARACTERIZATION OF NOVEL CORONAVIRUS GENOMES

This Chapter presents the quick and efficient development of one genotyping tools for Coronavirus and SARS-CoV-2. The tools method of the tools is to genotype by phylogenetic inference for this they do the aligning the user-submitted sequence with a carefully selected set of predefined reference strains, followed by Neighbour Joining (NJ) phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequences is assigned the genotype of the reference strain with the highest bootstrap scores (>70%). The tools allow high-throughput classification of these virus species and genotypes in seconds by providing users with a classified genotype report along with phylogenetic trees of the submitted sequences.

The Chapter is presented in the form of a research article. It has been published in the journal, *Bioinformatics*. The published manuscript is hereby presented in the journal format.

Manuscript Published: Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics. 2020 Jun 1;36(11):3552–5

Genome analysis

Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes

Sara Cleemput¹, Wim Dumon¹, Vagner Fonseca^{2,3,4}, Wasim Abdool Karim², Marta Giovanetti⁵, Luiz Carlos Alcantara^{3,5}, Koen Deforche^{1,*} and Tulio de Oliveira^{2,6,7,*}

¹Emweb bv, Herent, Belgium, ²KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa, ³Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, ⁴Coordenação Geral dos Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, (CGLAB/SVS-MS), Brasília, Brazil, ⁵Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil, ⁶Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa and ⁷Department of Global Health, University of Washington, Seattle, WA, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on January 31, 2020; revised on February 21, 2020; editorial decision on February 21, 2020; accepted on February 25, 2020

Abstract

Summary: Genome detective is a web-based, user-friendly software application to quickly and accurately assemble all known virus genomes from next-generation sequencing datasets. This application allows the identification of phylogenetic clusters and genotypes from assembled genomes in FASTA format. Since its release in 2019, we have produced a number of typing tools for emergent viruses that have caused large outbreaks, such as Zika and Yellow Fever Virus in Brazil. Here, we present the Genome Detective Coronavirus Typing Tool that can accurately identify the novel severe acute respiratory syndrome (SARS)-related coronavirus (SARS-CoV-2) sequences isolated in China and around the world. The tool can accept up to 2000 sequences per submission and the analysis of a new whole-genome sequence will take approximately 1 min. The tool has been tested and validated with hundreds of whole genomes from 10 coronavirus species, and correctly classified all of the SARS-related coronavirus (SARSr-CoV) and all of the available public data for SARS-CoV-2. The tool also allows tracking of new viral mutations as the outbreak expands globally, which may help to accelerate the development of novel diagnostics, drugs and vaccines to stop the COVID-19 disease.

Availability and implementation: <https://www.genomedetective.com/app/typingtool/cov>

Contact: koen@emweb.be or deoliveira@ukzn.ac.za

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

We are currently faced with a potential global epidemic of a new coronavirus that has infected thousands of people in China and is spreading rapidly around the world. In the end of January 2020, the WHO has declared it a global emergency (WHO, 2020). The novel coronavirus (SARS-CoV-2), first isolated in Wuhan, China, has already caused more infections than the previous severe acute respiratory syndrome (SARS) outbreak of 2002 and 2003. The virus is a SARS-related coronavirus (SARSr-CoV), and it is genetically

associated with SARSr-CoV strains that infect bats in China (Lu *et al.*, 2020; Zhu *et al.*, 2020). It causes severe respiratory illness, which the WHO recently named COVID-19 disease. It has high fatality rate (Huang *et al.*, 2020), can be transmitted from person to person, has infected over 70 000 individuals and has spread to over 30 countries in less than 2 months (WHO, 2020).

This coronavirus outbreak has been unprecedented; so too is the way that the scientific community has responded to it. They have openly and rapidly shared genomic and clinical data as never seen

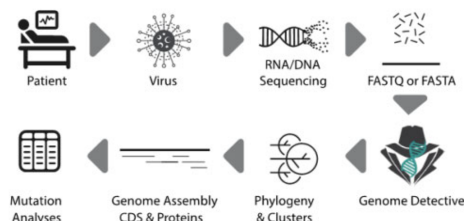


Fig. 1. Genome Detective Coronavirus Typing Tool assembles genomes from next-generation sequencing (NGS) in FASTAQ format or assembled genomes in FASTA format. A user can submit up to 1 Gb of NGS data or 2000 assembled genomic sequences. For each assembled genomic sequence, the tool identifies the virus species, constructs a phylogenetic tree and identifies phylogenetic clusters, which includes the novel coronavirus identified in Wuhan, China in 2019 (SARS-CoV-2). The tool identifies changes at nucleotides, coding regions and proteins using a novel dynamic aligner and display all of the mutations in detailed tables and reports

before allowing research results to be released almost instantaneously. This has helped the understanding of the transmission dynamics, the development of rapid diagnostic and has informed public health response. Here, we present a new contribution that can speed up this communal effort. The Genome Detective Coronavirus Typing Tool is a free-of-charge web-based bioinformatics pipeline that can accurately and quickly identify, assemble and classify coronavirus genomes. The tool also identifies changes at nucleotides, coding regions and proteins using a novel dynamic aligner to allow tracking new viral mutations (Fig. 1).

2 Systems and methods

A reference dataset of previously published coronavirus whole-genome sequences (WGS) was compiled from the Virus Pathogen Resource (VIPR) database (www.viprbrc.org). This dataset consisted of 386 WGS of nine important coronavirus species. These included 132 sequences of Severe Acute Respiratory Syndrome related Coronavirus (SARSr-CoV), 121 sequences of Beta coronavirus, 97 sequences of Middle East Respiratory Syndrome related Coronavirus (MERSr-CoV), 19 sequences of Human Coronavirus HKU1, 9 sequences of Murine Hepatitis Virus, 4 of Roussetus Bat Coronavirus HKU9, 3 of Rat Coronavirus and 1 WGS of Tylonycteris Bat Coronavirus HKU4, Zaria_bat_coronavirus and Longquan RI Rat coronavirus. To this reference dataset, we added 47 whole genomes of the current Coronavirus 2019 (SARS-CoV-2) outbreak that originated in Wuhan, China, in December 2019. The SARS-CoV-2 sequences were downloaded from the GISAID database (<https://www.gisaid.org>) together with annotation of its original location, collection date and originating and submitting laboratory. The SARS-CoV-2 data generators are properly acknowledged in the acknowledgements section of this article and detailed information is provided in Supplementary Table S1.

The 431 reference WGS were aligned with MUSCLE (Edgar, 2004). The alignment was manually edited until a codon alignment was attained in all coding sequences (CDS). A maximum likelihood phylogenetic tree, 1000 bootstrap replicates were constructed in PhyML (Guindon and Gascuel, 2003; Lemoine *et al.*, 2018) and a Bayesian tree using MrBayes (Ronquist and Huelsenbeck, 2003) were constructed. The trees were visualized in Figtree (Rambaut, 2018). We selected 25 reference sequences that represent the diversity of each well-defined phylogenetic cluster (with bootstrap support of 100% and posterior probability of 1). We identified five well-supported phylogenetic clusters with more than two sequences of SARSr-CoV and used them to set up our automated phylogenetic classification tool. Cluster 1 included SARS strains from the 2002 and 2003 Asian outbreaks. In our tool, we named this cluster *SARS-CoV Outbreak 2000s but may rename it as SARS-CoV-1 if a new proposed naming system for SARSr-Cov is adopted in the near*

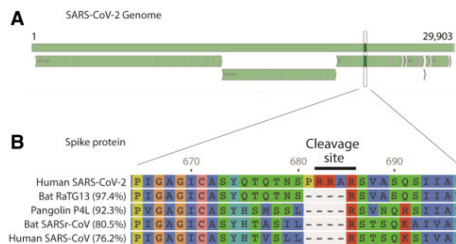


Fig. 2. Output from Genome Detective Coronavirus Typing Tool showing: (A) SARS-CoV-2 complete genome map. Top bar represents the genome (nucleotide positions 1 to 29 903). Bottom segments represent the open reading frames (ORFs). (B) Amino-acid alignment of the spike protein highlighting a four amino acid insertion (PRRA), which creates a new polybase cleavage site (RRAR) for SARS-CoV-2. Amino acid (aa) alignment is compared with four-related coronaviruses. The tool also calculates the percentage aa identities with reference to SARS-CoV-2 as shown here for the complete (1274 aa) spike protein

future. Cluster 2 (provisionally named as *SARS related CoV*) includes seven sequences from bats which did not cause large human outbreaks. Cluster 3 (named as *Bat SARS-CoV HKU3*) includes three WGS sampled from *Rhinolophus sinicus* (i.e. Chinese rufous horseshoe bats). Cluster 4 (*Bat SARS-CoV ZXC21/ZC45*) includes two SARSr-CoV sampled from *Rhinolophus sinicus* bats in Zhoushan, China. Cluster 5 (virus named *SARS-CoV-2 by the ICTV committee and disease named COVID-19 by the WHO*) includes three public sequences from the outbreak. We identified this cluster with many sequences from GISAID but kept only three ones as these were the first GenBank sequences. The first whole genome of SARS-CoV-2 was kindly shared by Prof. Yong-Zhen Zhang and colleagues in the virological.org website. Detailed information about the phylogenetic reference datasets is available in Supplementary Table S2.

The phylogenetic reference dataset was used to create an automated Coronavirus Typing Tool using the Genome Detective framework (Fonseca *et al.*, 2019; Vilsker *et al.*, 2019). To determine the accuracy of this tool, each of the 431 test WGS was considered for evaluation (i.e. 384 reference sequences from VIPR and 47 public SARS-CoV-2 sequences). The sensitivity, specificity and accuracy of our method were calculated for both species assignment and phylogenetic clustering of SARSr-CoV. Sensitivity was computed by the formula $\frac{TP}{TP+FN}$, specificity by $\frac{TN}{TN+FP}$ and accuracy by $\frac{TP+TN}{TP+FP+FN+TN}$, where TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives.

Classifying query sequences in an automated fashion involves two steps. The first step enables virus species assignments and the second, which is restricted to SARSr-CoV, includes phylogenetic analysis. The first classification analysis subjects a query sequence to BLAST and AGA analysis. AGA is a novel alignment method for nucleic acid sequences against annotated genomes from NCBI RefSeq Virus Database. AGA (Deforche, 2017) expands the optimal alignment algorithms of Smith and Waterman (1981) and Gotoh (1982) based on an induction state with additional parameters. The result is a more accurate aligner, as it takes into account both nucleotide and protein scores and identifies all of the polymorphisms at nucleotide and amino acid levels. In the second step, a query sequence is aligned against the phylogenetic reference dataset using -add alignment option in the MAFFT software (Katoh and Standley, 2013). In addition, a Neighbor-Joining phylogenetic tree is constructed using the HKY distance metric with gamma among-site rate variation with 1000 bootstrap replicates using PAUP* (Swofford, 2003). The query sequence is assigned to a particular phylogenetic cluster if it clusters monophyletically with that clade or a subset of it with bootstrap support >70%. If the bootstrap support is <70%, the genotype is reported to be unassigned.

The result of the phylogenetic and mutational analysis performed by AGA is available in a detailed report. This report

contains an interactive phylogenetic tree and genome mapper (Supplementary Fig. S1). It also presents the virus species and cluster assignments and a detailed table that provides information about open reading frames (ORFs), CDS and proteins. This table can be expanded to show nucleotide and amino acid mutations that differentiate a query sequence from their species RefSeq or from a sequence in the phylogenetic reference dataset. All results can be exported to a variety of file formats (XML, CSV, Excel, Nexus or FASTA).

3 Testing and validation

The Genome Detective Coronavirus Typing Tool correctly classified all of the 175 SARS-CoV sequences at species level, i.e. specificity, sensitivity and accuracy of 100%. Furthermore, all of the 47 SARS-CoV-2 WGS that were isolated in China ($n=36$), USA ($n=5$), France ($n=2$), Thailand ($n=2$), Japan ($n=1$) and Taiwan ($n=1$) were correctly classified at phylogenetic cluster level as SARS-CoV-2, which may be renamed as SARS-B. In addition, we classified with very high specificity, sensitivity and accuracy (i.e. 100%) all of the 112 SARS outbreak WGS of 2002 and 2003. We also achieved perfect classification (i.e. specificity, sensitivity and accuracy of 100%) for all of Beta coronavirus, Human coronavirus_HKU1, MERS-CoV, Roussetus Bat coronavirus HKU9 and Tylonycteris bat coronavirus HKU4 at species level. For a detailed overview of assignment performance, please refer to the Supplementary Table S3.

Our tool also allows detailed analysis of coding regions and proteins for each of the coronavirus species. For example, the analysis of the first released SARS-CoV-2 sequence, the WH_Human1_China_2019Dec (GenBank: MN908947) demonstrated at genome level, the nucleotide (NT) identity was 79.0% to the reference strain of SARS-CoV (ACCESSION: NC_004718.3) and that the Envelop Small Membrane Protein (protein E) is the most similar protein. In total, 94.8% (73/77) of the amino acids were identical; the four amino acid differences were located at positions 55 (T55S), 56 (V56F), 69 (69deletion) and 70 (G70R). The spike protein (protein S), which can be associated with virulence, was 76.2% identical to the reference strain of SARS-CoV (Supplementary Table S4A). Interestingly, there were four amino acid insertions at position 237 (A237_F238insHRSY; genome NT position 22202_22203insCATAGAAGTTAT), which is just upstream from a cleavage site. There is also a four amino acid insertion PRRA at the spike protein at positions 681 to 684. This is at the junction of S1 and S2 and creates a new polybase cleavage site. Our tool also allows us to compare mutations with other-related sequences, such as the Pangolin, Bat RaTG13, the Bat SARS-CoV and SARS Sin940 (Figure 2 and Supplementary Table S2). The most diverse coding regions were the CDS Sars8a and Sars8b. In these two regions, only 30% of the amino acids were identical. Sars8b protein was truncated early and its CDS had four stop codons (Supplementary Table S4A).

Our Coronavirus Typing Tool also allows a query sequence to be analyzed against a sequence in the phylogenetic reference dataset. For example, the WH_Human1_China_2019Dec (GenBank: MN908947) the identity was 87.5% to the Bat sequence bat_SL_CoVZXC21 (Genbank: MG772934). This was one of the Bat-CoV sequences that were most related to n2019-CoV (Lu et al., 2020). The Envelop Small Membrane Protein (protein E) was 100% identical (Supplementary Table S4B). When the SARS-CoV-2 isolated from France (BetaCoV/France/IDF0373/2020) was analyzed with our tool and compared with the SARS-CoV-2 WH_Human1_China_2019Dec strain (Accession: MN908947), this sequence was 99.9% identical and had only two NT mutations (Supplementary Table S4C). These two differences were located on positions: 22551G>T and 26016G>T, which caused three amino acid mutations (E2 glycoprotein Protein mutation: V354F (22551G>T), sars3a protein mutations: G250V (26016G>T) and sars3b protein mutations: V110F (26016G>T) (detailed in Supplementary Table

S4C-II). The analysis of a WGS in FASTA format takes approximately 60 s.

4 Discussion

We developed and released the Genome Detective Coronavirus Typing tool as a free-of-charge resource in the third week of January 2020 in order to help the rapid characterization of COVID-19 infections. This tool allows the analysis of whole or partial viral genomes within minutes. It accepts assembled genomes in FASTA format or raw next-generation sequencing data in FASTQ format from Illumina, Ion Torrent, PACBIO or Oxford Nanopore Technologies (ONT) can be submitted to the Genome Detective Virus Tool (Vilsker et al., 2019) to automatically assemble the consensus genome prior to executing the Coronavirus Typing Tool. User effort is minimal, and a user can submit multiple FASTA sequences at once.

The tool uses a novel and dynamic aligner, AGA, to allow submitted sequences to be queried against reference genomes, using both nucleotide and amino acid similarity scores. This allows accurate identification of other coronavirus species and the tracking of new viral mutations as the outbreak expands globally. It also performs detailed analysis of the coding regions and proteins. Moreover, it can easily be updated to add new phylogenetic clusters if new outbreaks arise or if the classification nomenclature changes. The tool has been able to correctly classify all the recently released SARS-CoV-2 genomes, as well as all the 2002–2003 SARS outbreak sequences.

In conclusion, the Genome Detective Coronavirus Typing Tool is a web-based and user-friendly software application that allows the identification and characterization of novel coronavirus genomes.

Acknowledgements

Genome data for SARS-CoV-2 made kindly available by National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC), Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Hubei Provincial Center for Disease Control and Prevention, Wuhan Institute of Virology, Chinese Academy of Sciences, National Institute for Viral Disease Control and Prevention, China CDC, Li Ka Shing Faculty of Medicine, the University of Hong Kong, Department of Medical Sciences, Ministry of Public Health, Thailand | Thai Red Cross Emerging Infectious Diseases—Health Science Centre | Department of Disease Control, Ministry of Public Health, Thailand, Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention, Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Division of Viral Diseases, Centers for Disease Control and Prevention, Centers for Disease Control, R.O.C. (Taiwan) | Centers for Disease Control, R.O.C. (Taiwan), California Department of Public Health | Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, Arizona Department of Health Services | Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, Guangdong Provincial Centers for Diseases Control and Prevention | Guangdong Provincial Institute of Public Health, University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong, State Key Laboratory of Virology, Wuhan University, Department of Infectious and Tropical Diseases, Bichat Claude Bernard Hospital, Paris | National Reference Center for Viruses of Respiratory Infections, Institute Pasteur, Paris. We would like to acknowledge all of the data contributors.

Funding

Research reported in this publication was supported by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI), the VIROGENESIS project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme (under Grant Agreement no. 634650) and the National Human

Genome Research Institute of the National Institutes of Health under Award Number U24HG006941. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: Emweb bv is the company that makes the Genome Detective Coronavirus Typing Tool online available. S.C. is employee of emweb bv. K.D. and W.D. are directors and owners of emweb bv. This company has allowed the coronavirus typing tool to freely available on the web.

References

- Deforche, K. (2017) An alignment method for nucleic acid sequences against annotated genomes. *Biorxiv*. doi:10.1101/200394.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
- Fonseca, V. *et al.* (2019) A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes. *PLoS Negl. Trop. Dis.*, **13**, e0007231.
- Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
- Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
- Huang, C. *et al.* (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, **395**, 497–506.
- Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
- Lemoine, F. *et al.* (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, **556**, 452–456.
- Lu, R. *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, **395**, 565–574.
- Rambaut, A. (2018) FigTree v1.4.4. *Institute of Evolutionary Biology*. University of Edinburgh, Edinburgh. <http://tree.bio.ed.ac.uk/software/figtree/> (31 January 2020, date last accessed).
- Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Swofford, D.L. (2003) *PAUP* 4.0: Phylogenetic Analysis Under Parsimony (And Other Methods), Version 4.0b2a*. Sinauer Associates Inc., Sunderland, MA.
- Vilsker, M. *et al.* (2019) Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics*, **35**, 871–873.
- World Health Organization (WHO) (2020) Novel Coronavirus (COVID-19) Situational Reports. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/> (17 February 2020, date last accessed).
- Zhu, N. *et al.*; China Novel Coronavirus Investigating and Research Team (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*, **382**, 727–733.

CHAPTER 5: WEST NILE VIRUS IN BRAZIL

This Chapter presents the quick and efficient development of one genotyping tools for West Nile Virus. The tools method of the tools is to genotype by phylogenetic inference for this they do the aligning the user-submitted sequence with a carefully selected set of predefined reference strains, followed by Neighbour Joining (NJ) phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequences is assigned the genotype of the reference strain with the highest bootstrap scores (>70%). The tools allow high-throughput classification of these virus species and genotypes in seconds by providing users with a classified genotype report along with phylogenetic trees of the submitted sequences.

The Chapter is presented in the form of a research article. It has been published in the journal, *Pathogens*. The published manuscript is hereby presented in the journal format.

Manuscript Published: Costa ÉA, Giovanetti M, Silva Catenacci L, Fonseca V, Aburjaile FF, Chalhoub FLL, et al. West nile virus in brazil. Pathogens. 2021 Jul 15;10(7).

Article

West Nile Virus in Brazil

Érica Azevedo Costa ^{1,†}, Marta Giovanetti ^{2,3,†}, Lílilã Silva Catenacci ^{4,†}, Vagner Fonseca ^{3,5,6,†}, Flávia Figueira Aburjaile ^{3,†}, Flávia L. L. Chalhoub ², Joilson Xavier ³, Felipe Campos de Melo Iani ⁷, Marcelo Adriano da Cunha e Silva Vieira ⁸, Danielle Freitas Henriques ⁹, Daniele Barbosa de Almeida Medeiros ⁹, Maria Isabel Maldonado Coelho Guedes ¹, Beatriz Senra Álvares da Silva Santos ¹, Aila Solimar Gonçalves Silva ¹, Renata de Pino Albuquerque Maranhão ¹⁰, Nieli Rodrigues da Costa Faria ², Renata Farinelli de Siqueira ¹¹, Tulio de Oliveira ⁵, Karina Ribeiro Leite Jardim Cavalcante ¹², Noely Fabiana Oliveira de Moura ¹², Alessandro Pecego Martins Romano ¹², Carlos F. Campelo de Albuquerque ¹³, Lauro César Soares Feitosa ¹⁴, José Joffre Martins Bayeux ¹⁵, Raffaella Bertoni Cavalcanti Teixeira ¹⁶, Osmakon Lisboa Lobato ¹⁷, Silvokleio da Costa Silva ¹⁷, Ana Maria Bispo de Filippis ², Rivaldo Venâncio da Cunha ¹⁸, José Lourenço ¹⁹ and Luiz Carlos Junior Alcantara ^{2,3,*}



Citation: Costa, É.A.; Giovanetti, M.; Silva Catenacci, L.; Fonseca, V.; Aburjaile, F.F.; Chalhoub, F.L.L.; Xavier, J.; Campos de Melo Iani, F.; da Cunha e Silva Vieira, M.A.; Freitas Henriques, D.; et al. West Nile Virus in Brazil. *Pathogens* **2021**, *10*, 896. <https://doi.org/10.3390/pathogens10070896>

Academic Editor: Francisco Llorente

Received: 30 April 2021

Accepted: 21 May 2021

Published: 15 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

- ¹ Departamento de Medicina Veterinária Preventiva, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil; azevedoc@yaho.com.br (É.A.C.); mariaisabel.guedes@gmail.com (M.I.M.C.G.); beatrizsenra.santos@gmail.com (B.S.Á.d.S.S.); ailavet@yahoo.com.br (A.S.G.S.)
- ² Laboratório de Flavivirus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro 21040-360, Brazil; giovanetti.marta@gmail.com (M.G.); flaviallevy@yahoo.com.br (F.L.L.C.); nielircf@gmail.com (N.R.d.C.F.); ana.bispo@ioc.fiocruz.br (A.M.B.d.F.)
- ³ Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil; vagner.fonseca@gmail.com (V.F.); faburjaile@gmail.com (F.F.A.); joilsonxavier@live.com (J.X.)
- ⁴ Departamento De Morfofisiologia Veterinária, Universidade Federal do Piauí, Teresina 64049-550, Brazil; catenacci@ufpi.edu.br
- ⁵ KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban 4041, South Africa; tulioDNA@gmail.com
- ⁶ Coordenação Geral dos Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde (CGLAB/SVS-MS), Brasília 70719-040, Brazil
- ⁷ Laboratório Central de Saúde Pública, Fundação Ezequiel Dias, Belo Horizonte 30510-010, Brazil; felipeemrede@gmail.com
- ⁸ Diretoria de Vigilância em Saúde, Fundação Municipal de Saúde, Teresina 64600-000, Brazil; marceloadrianoneuro@gmail.com
- ⁹ Seção de Arbovirologia e Febres Hemorrágicas, Instituto Evandro Chagas, Ministério da Saúde, Ananindeua 70058-900, Brazil; dannifh@hotmail.com (D.F.H.); danielmedeiros@iec.pa.gov.br (D.B.d.A.M.)
- ¹⁰ Setor de Clínica de Equinos, Hospital Veterinário, Campus Pampulha, Universidade Federal de Minas Gerais Escola de Veterinária, Belo Horizonte 31270-901, Brazil; rpamaranhao@yahoo.com
- ¹¹ Department of Large Animal Clinic, Universidade Federal de Santa Maria, Rio Grande do Sul 97105-900, Brazil; renata.farinelli@ufsm.br
- ¹² Coordenação Geral das Arboviroses, Secretaria de Vigilância em Saúde/Ministério da Saúde, Brasília 70058-900, Brazil; karina.cavalcante@saude.gov.br (K.R.L.J.C.); noely.moura@saude.gov.br (N.F.O.d.M.); alessandro.romano@saude.gov.br (A.P.M.R.)
- ¹³ Organização Pan-Americana da Saúde, Organização Mundial da Saúde, Brasília 40010-010, Brazil; meloc@paho.org
- ¹⁴ Centro de Ciências Agrárias, Departamento de Clínica e Cirurgia Veterinária, Universidade Federal do Piauí, Teresina 64049-550, Brazil; jackvet08@hotmail.com
- ¹⁵ Faculdade de Ciências da Saúde, Medicina Veterinária, Urbanova, São José Dos Campos, UNIVAP-Universidade Vale do Paraíba, São Paulo 12245-720, Brazil; jveterinario@hotmail.com
- ¹⁶ Departamento de Clínica e Cirurgia Veterinárias, Escola de Veterinária, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil; teixeiraraffa@gmail.com
- ¹⁷ Laboratório de Genética e Conservação de Germoplasma, Campus Prof. Cinobelina Elvas, Universidade Federal do Piauí, Bom Jesus, Piauí 64049-550, Brazil; osmaikonlobato@gmail.com (O.L.L.); silvokleio@ufpi.edu.br (S.d.C.S.)
- ¹⁸ Coordenação dos Laboratórios de Referência, Oswaldo Cruz Foundation, Rio de Janeiro 21040-360, Brazil; rivaldo.cunha@ioc.fiocruz.br
- ¹⁹ Department of Zoology, University of Oxford, Oxford OX1 3PS, UK; jose.lourenco@zoo.ox.ac.uk
- * Correspondence: luiz.alcantara@ioc.fiocruz.br

† Denote equal contribution.

Abstract: *Background:* West Nile virus (WNV) was first sequenced in Brazil in 2019, when it was isolated from a horse in the Espírito Santo state. Despite multiple studies reporting serological evidence suggestive of past circulation since 2004, WNV remains a low priority for surveillance and public health, such that much is still unknown about its genomic diversity, evolution, and transmission in the country. *Methods:* A combination of diagnostic assays, nanopore sequencing, phylogenetic inference, and epidemiological modeling are here used to provide a holistic overview of what is known about WNV in Brazil. *Results:* We report new genetic evidence of WNV circulation in southern (Minas Gerais, São Paulo) and northeastern (Piauí) states isolated from equine red blood cells. A novel, climate-informed theoretical perspective of the potential transmission of WNV across the country highlights the state of Piauí as particularly relevant for WNV epidemiology in Brazil, although it does not reject possible circulation in other states. *Conclusion:* Our output demonstrates the scarceness of existing data, and that although there is sufficient evidence for the circulation and persistence of the virus, much is still unknown on its local evolution, epidemiology, and activity. We advocate for a shift to active surveillance, to ensure adequate preparedness for future epidemics with spill-over potential to humans.

Keywords: West Nile virus; genomic monitoring; molecular detection; Brazil

1. Introduction

West Nile virus (WNV), a member of the *Flaviviridae* family, was first identified in the West Nile district of Uganda in 1937, but nowadays, it is commonly found in Africa, Europe, North America, the Middle East, and Asia [1–3]. WNV transmission is maintained in a mosquito–bird cycle, for which the genus *Culex*, in particular *Cx. pipiens* and *quinquefasciatus*, are considered the principal vectors [4]. WNV can infect humans, equines, and other mammals, but these are considered “dead-end” hosts, given their weak potential to function as amplifying hosts to spread infection onwards [5,6]. Around 80% of WNV infections in humans are asymptomatic, while the rest may develop mild or severe disease. Mild disease includes fever, headache, tiredness, and vomiting [7,8], while severe disease (neuroinvasive) is characterized by high fever, coma, convulsions, and paralysis [7,8]. Equine infections can occasionally cause neurological disease and death [7,8], such that equines typically serve as sentinel species for WNV outbreaks with potential for spill-over into human populations.

Genome detection of WNV in South America was originally reported in horses (Argentina in 2006) and captive flamingos (Colombia, in 2012) [9,10]. The first ever sequenced genome in Brazil was in 2018, when the virus was isolated from a horse with severe neurological disease in the Espírito Santo state [11]. Despite multiple studies reporting serological evidence suggestive of past WNV circulation in Brazil (e.g., [11–13]) and reports of human WNV disease in confirmed cases in the Piauí state [13], much is unknown about genomic diversity, evolution, and transmission dynamics across the country. The reality of WNV in Brazil is likely characterised by endemic circulation within the mosquito–bird cycle [14–17], with occasional transmission to humans. The so far lack of reported human epidemics with significant public health impact remains a puzzle, given that Brazil harbors the necessary vectors, avian species, and climate—combination amenable at sustaining endemicity [18]. Several factors potentially contribute to the seemingly silent circulation of WNV in the country [19], such as the lack of surveillance interest and resources, rates of mild human WNV disease, co-circulation of other mosquito-borne viruses that cause similar clinical spectrums, and diagnostics and screening of animals and humans well past the time of infection, which critically hampers viral detection and confirmation.

In this study, we aim at providing a holistic perspective of what is known about WNV circulation in Brazil. In addition to previously reported evidence of WNV circulation, we

also report new genetic evidence of WNV circulation in three Brazilian states. We further provide a climate-informed, theoretical assessment of the transmission potential of WNV across Brazil, revealing spatio-temporal patterns of interest. The lack of surveillance data hampers more in-depth analyses and therefore obscures our current understanding of WNV epidemiology, evolution, and transmission in the country. Recently, some European countries have witnessed a shift from a similar surveillance and epidemiological situation to that of Brazil, to observing recurrent WNV epidemics with spill-over to human populations [18–21]. We argue that active surveillance initiatives are necessary in Brazil in the near future to ensure preparedness of future WNV epidemics with public health impact.

2. Results

2.1. Novel Evidence of WNV Circulation in Three Brazilian States

Samples (RBCs) from three horses with suspected WNV infection obtained from southern (Minas Gerais and São Paulo) and northeastern (Piauí) Brazilian states were sent for molecular diagnosis at the Departamento de Medicina Veterinária Preventiva at the Federal University of Minas Gerais (UFMG).

RNAs were extracted from red blood cells and tested using an in-house PCR assay (see Methods section for details). WNV-specific RT-PCR amplification products were obtained by nested PCR (Figure 1A,B), and positive samples were subjected to a newly designer multiplex PCR scheme (Supplementary Table S1) to generate complete genomes sequences by means of portable nanopore sequencing.

Three blood fractions (plasma, buffy coat, and washed RBC) from the horses sampled in São Paulo and Minas Gerais states have been submitted to nested RT-PCR; horse samples from Piauí have been tested only using RBC, which was the only blood fraction available. Diagnostic investigation of alphavirus was also performed using a generic RT-PCR targeting the NSP1 gene, according to [22], in the three blood fractions, with negative results.

The published WNV genome from Brazil (MH643887) was used to generate (mean) 98.4% consensus sequences that formed the target for primer design. The new genomes were deposited in GenBank with accession numbers MW420987, MW420988, and MW420989 (Table 1).

We constructed phylogenetic trees to explore the relationship of the sequenced genomes to those sampled elsewhere globally. We retrieved 2321 WNV genome sequences with associated lineage date and country of collection from GenBank, from which we generated a subset that included the highly supported (>0.9) clade containing the newly WNV strains obtained in this study plus 29 sequences (randomly sampled) from all lineages and performed phylogenetic analysis. An automated online phylogenetic tool to identify and classify WNV sequences was developed (available at: <http://krisp.ukzn.ac.za/app/typingtool/wnv/job/9b40f631-51c4-419c-9edf-2206e7cd8d9c/interactive-tree/phylo-WNV.xml> accessed on 31 December 2019).

Phylogenies estimated by the newly developed WNV typing tool, along with maximum likelihood methods (Supplementary Figure S1C), consistently placed the Brazilian genomes in a single clade within the 1a lineage with maximum statistical support (bootstrap = 100%) (Supplementary Figure S1).

Time-resolved maximum likelihood tree appeared to be consistent with previous estimates [11] and showed that the new genomes clustered with strong bootstrap support (97%) with a WNV strain isolated from an *Aedes albopictus* mosquito in Washington DC, USA in 2019 (Figure 1D). Interestingly, the new isolates did not group with the previously sequenced genome in 2019 from the Espírito Santo state, suggesting that inter-continental introduction events might be frequent in Brazil.

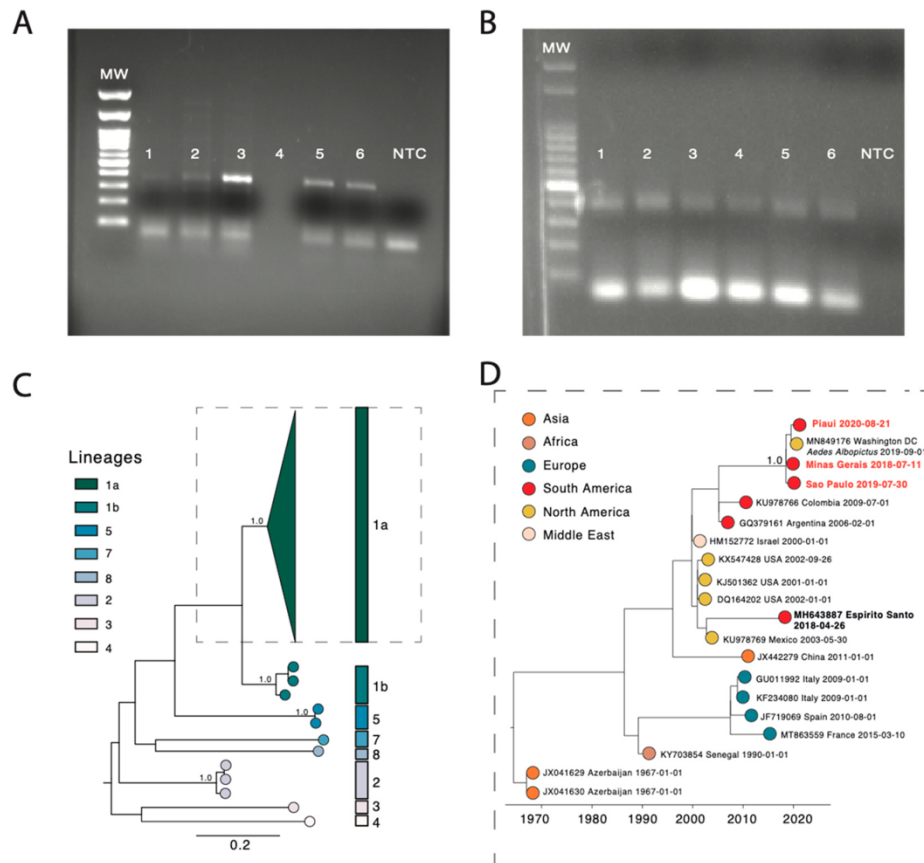


Figure 1. Investigation of WNV infections in Brazil, between July 2018 and September 2020, and estimated transmission potential. **(A,B)** Agarose gel electrophoresis of amplicons from assay for WNV. **(A)** nested RT-PCR. MW (Molecular weight ladder), 100 bp DNA Ladder RTU, Kasvi; 1—plasma of horse from São Paulo; 2—buffy coat of horse from São Paulo; 3—washed RBC of horse from São Paulo; 4—blank negative control using during the nested RT-PCR; 5 and 6—positive control (synthetic gene); NTC, no template control (using since the extraction); expected amplicon size: 370 bp. **(B)** Multiplex PCR. MW (Molecular weight ladder), Fluorescent 100 bp DNA Ladder, Cellco, Jena Bioscience; 1—horse form Minas Gerais (pair primers); 2—horse form Minas Gerais (odd primers); 3—horse form Sao Paulo (pair primers); 4—horse form Sao Paulo (impair primers); 5—horse form Piaui (pair primers); 6—horse form Piaui (odd primers); NTC, no template control (using since the extraction); expected amplicon size: 400 bp. **(C)** Midpoint rooted maximum-likelihood phylogeny of WNV genomes, showing major lineages. The scale bar is in units of substitutions per site (s/s). Support for branching structure is shown by bootstrap values at nodes. **(D)** Time-resolved maximum likelihood tree showing the WNV strains belonged to the 1a lineage. Colors indicate geographic location of sampling. The new Brazilian WNV strains are shown with text in red.

Table 1. Epidemiological information and sequencing statistics of the three sequenced samples of WNV sampled in Minas Gerais, São Paulo, and Piauí Brazilian states.

ID	Sample	Collection Date	Age	Sex	State	Municipality	Reads	Coverage (%)	Depth of Coverage	Lineage Assignment	Accession Number	Clinical Sign
BC02_07	RBCs	11/07/2018	9 months	F	MG	Sabara	343,743	97.9	6527.6	Lineage 1a	MW420989	Chorioretinitis
BC03_04	RBCs	30/07/2019	13 years-old	M	SP	São Bernardo do Campo	170,980	97.9	3189.7	Lineage 1a	MW420988	Muscle stiffness, tremor retinal and flaccid paralysis
BC05_06	RBCs	21/08/2020	5 years-old	F	PI	Parnaíba	222,516	99.4	4121.4	Lineage 1a	MW420987	Neurological complications

ID = study identifier; RBCs = Red Blood Cells; Collection date = Sample collection date; Municipality = Municipality of residence; State = MG-Minas Gerais; SP = Sao Paulo; PI = Piauí; Sex: M = Male; F = Female; Accession Number = NCBI accession number.

2.2. A Data-Driven WNV Theoretical Perspective

We first summarized the past evidence of WNV circulation in Brazil from avian species, equines, and humans, which was achieved via various literature reports using different confirmation methods (Figure 2A) [23]. The first evidence of WNV infection was documented in 2004 in horses in northeastern Brazil (Paraíba state). Since then, serological evidence of WNV infection continued to be documented between 2008 and 2010 and again in 2020 in horses and birds from the southern [24], midwestern (Pantanal), and northern Brazilian regions. In 2014, the first WNV infection in a human was confirmed in the Piauí State (northeast region). In 2018, the first isolation of WNV in Brazil was documented in the Espírito Santo state (southeastern Brazil) when the virus was isolated from the central nervous system (CNS) of a dead horse with neurological manifestations [11]. To these data, we here add the report of the new genetic evidence of WNV circulation in equines occurring between 2018 and 2020, in southern (Minas Gerais, São Paulo) and northeastern (Piauí) states. To the best of our knowledge, it is the first time that evidence of WNV circulation is reported for the states of Minas Gerais and São Paulo.

Using data collected from the Brazilian “Sistema de Informação de Agravos de Notificação” (SINAN) (see Methods and Supplementary Table S2) reported with identifier A923 (“Febre do Nilo”), we explored the current spatio-temporal distribution of suspected cases of West Nile fever. Given the unspecific and unconfirmed nature of these reported cases, we complemented such information with theoretical projections of the spatio-temporal transmission potential of WNV in Brazil. For this, we used a climate-driven suitability measure (index P) previously successfully applied to WNV in the contexts of Israel [25] and Portugal [26] (see Methods).

We mapped the mean index P across Brazil for the period 2015–2019 (Figure 2B) and found estimated transmission potential to be highest in the center of the country along a diagonal latitude–longitude axis crossing from the center–west to the north–east. The regions of the south of the country, similarly to estimations for other mosquito-borne viruses [27], presented the least transmission potential. To assess potential hotspots of (at least temporary) high transmission potential, we calculated the proportion of months (2015–2019) in which the index P was above 1; this particular threshold representing the point above which each female mosquitoes would be theoretically able to infect more than one host during their lifetime. This approach identified regions of Piauí, Bahia, Ceará, Rio Grande do Norte, and Paraíba states as presenting significantly longer periods of time with high index P . In particular, the state of Piauí was captured in its entirety within this estimated spatial hotspot of transmission potential (Figure 2C).

From all states for which there were reported cases, we filtered those that had more than one case per any month during the entire period of 2015–2019, selecting only two states with clear epidemic waves of reported cases: Piauí and Espírito Santo. Coincidentally with the results of Figure 2B,C, the state of Piauí reported the largest number of cases in the entire dataset. Using the geographical boundaries of each state, we averaged the index P per month (Figure 2D,E). The resulting time series of transmission potential showed that potential was higher in Piauí compared to Espírito Santo in accordance with the spatial output in Figure 2B,C. It also presented a clear seasonal signal, with peaks occurring on average in February in Piauí (month average = 2.2, summer) and April in Espírito Santo (month average = 4, autumn). The correlation between reported cases and the index for Piauí was positive (Pearson’s 0.36, Figure 2D), but it was negative for Espírito Santo (Pearson’s -0.31 , Figure 2E). Similar to what has been reported for suitability indices applied to other viruses [27], there was a clear lag between the index and cases for Piauí, with cases lagging behind the index (Figure 2D). Accordingly, shifting the index by one month into the future resulted in a high positive correlation with cases (Pearson’s 0.84).

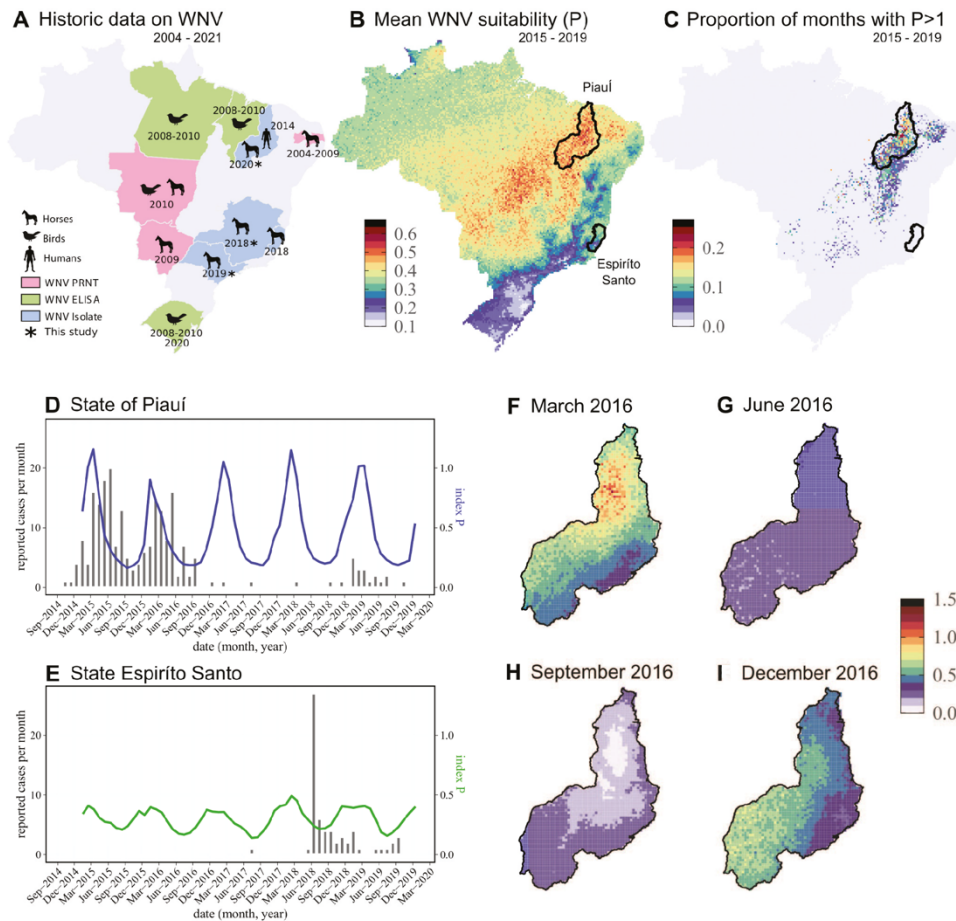


Figure 2. Data-driven epidemiological perspective of WNV in Brazil. **(A)** Mapping of historic evidence for WNV circulation in Brazil, for which the color and symbol legend on the bottom left of the panel define the animal source and methodology. Data are based on a literature review up to 2019 [24], in addition with recently published reports in 2020–2021 [24] and the new data generated in this study. **(B)** Mean estimated transmission potential of WNV (index P) over the period 2015–2019. The color scale on the bottom left of the panel shows the range of the presented values. The black borders mark the boundaries of the Piauí and Espírito Santo states. **(C)** Proportion of months for which the transmission potential of WNV (index P) was above the value 1, over the period 2015–2019. The color scale on the bottom left of the panel shows the range of the presented values. The black borders mark the boundaries of the Piauí and Espírito Santo states. **(D)** Time series of suspected reported West Nile fever cases (bars) and estimated transmission potential of WNV (index P , blue line) for the Piauí state. Index P is the average per month, across all data points within the boundaries of the state. **(E)** Time series of suspected reported West Nile fever cases (bars) and estimated transmission potential of WNV (index P , green line) for the Espírito Santo state. Index P is the average per month, across all data points within the boundaries of the state. **(F)** Spatial snapshot of estimated transmission potential of WNV (index P) for the month of March 2016. Color scale on the right shows the range of the presented values. **(G)** Same as F but for June 2016. **(H)** Same as F but for September 2016. **(I)** Same as F but for December 2016.

Finally, to get a grasp of the possible spatio-temporal dynamics of WNV transmission in Piauí, we looked at estimated transmission potential for one of the years with more reported suspected cases (2016) both in space and time (with snapshots at months of March, June, September, and December) (Figure 2F–I). The spatio-temporal snapshots showed that transmission potential was the lowest during winter months, but we also highlighted that this was almost uniform across the state (Figure 2G,H). In contrast, throughout the year, this output highlighted a possible wave of seasonal transmission. This wave would typically start in the southwest just before summer (Figure 2I) and would move to the northeast in the summer (Figure 2F).

3. Discussion

Our analyses indicate that additional data are required to better identify routes of WNV importation into and within Brazil and to more generally understand the local transmission dynamics of the virus. Interestingly, our data suggest that the circulation of the virus may have resulted from multiple independent introductions, since the new isolates did not group with the previously sequenced genome in 2019 from the Espírito Santo state. This suggests that intra-continental introduction events due to the mobility of infected birds or mosquitoes might be a more plausible mechanism for the multiple introductions of WNV in South American countries, including Brazil. This scenario is consistent with previous studies that showed that multiple independent introductions into Latin America occurred during the initial outbreak in US in 1999; detailed revision is provided in [28]. While migrating birds are a convenient explanation of WNV dispersal, other possible ways of dispersion exist, such as infected mosquitoes that are accidentally transported via airplane or by road transport [29]. Another likely scenario is commercial legal or illegal human transportation of birds and/or mosquitoes, which could be transported on airplanes [29].

The current data scarceness prevents definite conclusions on key aspects of WNV epidemiology. For example, given the unconfirmed nature of the reported cases by SINAN for Piauí and Espírito Santo, it is unclear what the proportion of cases truly reflect WNV occurrence and seasonality, hampering our ability to ascertain how representative our theoretical projections are. For Piauí, we would speculate that reported cases may indeed reflect some aspects of WNV seasonality, given that this state had the largest number of cases reported while also being the region of Brazil for which we estimated higher transmission potential and that our estimated transmission potential was well correlated with reported cases (albeit with a possible lag of one month typical of mosquito-borne viruses). At the same time, while inferred trees including the new genome sequences suggest that inter-continental introduction events might be frequent in Brazil, the lack of higher spatio-temporal sampling restricts our ability for definite conclusions on viral movement and persistence.

The phylogenetic and epidemiologic perspectives presented in this study, based on both existing and novel data as well as theoretical projections, suggest that both scenarios of sporadic and endemic local transmission are possible [30]. Similarly to sudden changes in WNV epidemiology and transmission as recently observed in other countries, the occurrence of a WNV outbreak affecting humans in Brazil may simply be a matter of time. Shifting from passive to active WNV screening and sequencing in animal reservoirs (e.g., equines, birds, vectors) in Brazil must be implemented to better understand the virus' local epidemiology and to be able to act accordingly in preventing and controlling any future epidemics with spill-over to humans.

4. Materials and Methods

4.1. Sample Collection, Viral RNA Isolation and PCR Screening

Samples (red blood cells, RBCs) from three horses with suspected WNV infection obtained from southern (Minas Gerais and São Paulo) and northeastern (Piauí) Brazilian states were sent for molecular diagnosis at the Laboratório de Patologia Molecular at the Federal University of Minas Gerais (UFMG).

Sample 1 from 11 July 2018 was collected from a 9-month-old female horse in a farm in the state of Minas Gerais, Mangueiras neighbourhood (Sabará), 15 km from the capital Belo Horizonte. Clinical findings were consistent with bilateral blindness. Neurological examination revealed no other abnormalities. The ophthalmological exams (direct and indirect pupillary light reflex (PLR), fluorescein eye stain test, fundus examination, and intraocular pressure) were consistent with retinal disease, mainly with chorioretinitis.

Sample 2 from 30 July 2019 was collected from a 13-year-old male horse that presented seizure episodes, muscle stiffness, tremor retinal, and flaccid paralysis in a farm located in São Bernardo do Campo countryside of the São Paulo state. Twenty-four days after the onset of neurological signs, the animal had severe pain in the forelimbs from laminitis, and it was euthanized due to hoof decumulation.

Sample 3 from 21 August 2020 was collected from a male horse, 5 years old, which died 72 h after presenting neurological signs, in a farm located in the municipality of Parnaíba, Piauí state. The animal presented motor incoordination, paddling movements, loss of sensitivity over the spine column, and behavioral changes. In this municipality, the tenth human case in Brazil was also detected, presenting neuroinvasive disease compatible with WNV infection, confirmed by serological assay (IgM) in both serum and cerebrospinal fluid (CSF) samples during acute and convalescent phases.

Whole blood samples obtained from the three horses were centrifuged at $1260\times g$ for 20 min, and the plasma and buffy coat fractions were collected and stored at $4\text{ }^{\circ}\text{C}$. Red blood cells (RBC) were washed by centrifugation three times in phosphate-buffered saline (PBS) at $1260\times g$ for 10 min and stored also at $4\text{ }^{\circ}\text{C}$ [15]. RNA from each unit (washed RBC, plasma and buffy coat) were extracted using the QIAmp Viral RNA Mini kit (Qiagen, Hilden, Germany), following manufacturer's recommendations.

Diagnostic investigation of arboviruses was performed by a generic RT-PCR targeting the flavivirus non-structural protein 5 (NS5) gene [31] and alphavirus non-structural protein 1 gene (nsP1) [32]. West Nile virus-specific degenerated primers: forward primers (+) AACCKCCAGAAGGAGTSAAR and reverse primers (−) AGCYTCRAACTCCAGRAAGC were used in second reaction of nested PCR targeting the NS5 gene after a genus specific flavivirus RT-PCR amplification [22]. A synthetic gene fragment of partial NS5 gene (gblocks gene fragment, Integrated DNA Technologies) was used as a positive control. The 25 μL PCR "master-mix" comprised 2.5 μL of $10\times$ PCR buffer, 1.5 mM MgCl_2 , 0.4 μM of each primer (forward and reverse), 0.8 μM dNTP mixture (Phonutria, Sao Paulo, Brazil), 1 U Taq DNA polymerase (Platinum Taq DNA polymerase; Invitrogen, Carlsbad, CA, USA), 2 μL of template DNA (sample or gBlock), and DNA/RNase-free water. The thermocycling conditions involved 40 cycles, and reaction conditions were previously reported in [18]. As an internal control for amplification efficiency, primers for the beta actin gene were used. As a negative control for the reactions, we used RNA extracted from equine washed RBC, plasma, and buffy coat that previously tested negative for arboviruses, equine herpesvirus 1 and 4, and borna disease. The amplicons were analyzed by 1% (*w/v*) agarose gel electrophoresis, stained with ethidium bromide, and visualized under UV light. Nested PCR were performed for equine herpesvirus 1 (EHV-1) [33] for borna disease [34,35], both with negative results in the 3 horses.

4.2. cDNA Synthesis and Multiplex Tiling PCR

Then, WNV-positive (in nested RT-PCR) RNA samples from washed RBCs were submitted to a cDNA synthesis protocol [36] using a Superscript IV cDNA Synthesis Kit. Then, a multiplex PCR primer scheme was designed (Table S1) to generate complete genomes sequences by means of portable nanopore sequencing, using Primal Scheme (Supplementary Table S1) (<http://primal.zibraproject.org> accessed on 31 December 2019) [37]. The published WNV genome from Brazil (MH643887) was used to generate a mean 98.4% consensus sequences that formed the target for primer design. The thermocycling conditions involved 40 cycles, and reaction conditions were previously reported in [37].

4.3. Library Preparation and Nanopore Sequencing

Amplicons were purified using 1× AMPure XP Beads, and cleaned-up PCR products concentrations were measured using Qubit™ dsDNA HS Assay Kit on a Qubit 3.0 fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA). DNA library preparation was carried out using the Ligation Sequencing Kit and the Native Barcoding Kit (NBD104, Oxford Nanopore Technologies, Oxford, UK) [37]. Purified PCR products pools were pooled together before barcoding reactions (taking in consideration each amplicon pool DNA concentrations), and one barcode was used per sample in order to maximize the number of samples per flow cell. Sequencing library was loaded onto a R9.4 flow cell, and data were collected for up to 6 h, but generally less.

4.4. Generation of Consensus Sequences

Raw files were basecalled using Guppy and barcode demultiplexing was performed using qcat. Consensus sequences were generated by de novo assembling using Genome Detective (<https://www.genomedetective.com/app/> accessed on 31 December 2019) [38]. New genomes were deposited in the GenBank with accession numbers MW420987, MW420988, and MW420989 (Table 1).

4.5. West Nile Virus Typing Tool: Classification Method and Implementation

The classification pipeline we present comprises two components. One for species and sub-species assignment that enables assignment at these levels by BLASTing the query sequences against a set reference sequences [39]. An assignment is made when BLAST reports a result that exceeds the present threshold.

The other component constructs a Neighbor Joining (NJ) phylogenetic tree that is used to make assignments at the lineages and sublineages level. For this component, the query sequence is aligned against a set of reference sequences using the profile alignment option in the ClustalW software [40], such that the query sequence is added to the existing alignment of reference sequences. Following the alignment, a NJ phylogenetic tree with 100 bootstrap replicates is inferred. The tree is constructed using the HKY distance metric with gamma among-site rate variation, as implemented in the PAUP* software (<https://paup.phylosolutions.com/> accessed on 31 December 2019) [41]. The query sequence is assigned to a particular genotype if it clusters monophyletically with that genotype clade with bootstrap support >70%. If the bootstrap support is <70%, the genotype is reported to be unassigned (Supplementary Figure S1).

For each of these steps, the earlier discussed reference strains were used with respect to the appropriate typing level (i.e., virus species, lineages, and sublineages). Testing revealed that a BLAST cut-off value of 200 allowed accurate identification of the virus species and WNV using sequence segments >200 base pairs. Note that the species classification procedure is implemented as separate BLAST steps. This enables the tool to efficiently perform large throughput species classification, such as for the classification of shorts sequencing reads. An instance of the web application is publically available on a dedicated server (<https://www.genomedetective.com/app/typingtool/wnv/> accessed on 31 December 2019). The web interface on this server accepts up to 2000 whole-genome or partial genome sequences at a time.

4.6. Phylogenetic Analysis

The 3 new sequences reported in this study were initially submitted to a genotyping analysis using the new phylogenetic West Nile virus subtyping tool, which is available at <https://www.genomedetective.com/app/typingtool/wnv/> (accessed on 31 December 2019). To put the newly WNV sequences in a global context, we constructed phylogenetic trees to explore the relationship of the sequenced genomes to those of other isolates.

We retrieved 2321 WNV genome sequences with associated lineage date and country of collection from GenBank (Supplementary Figure S2). From this dataset, we generated a subset that included the highly supported (>0.9) clade containing the newly WNV

strains obtained in this study plus 29 globally sequences (randomly sampled) from all lineages 1A, 1B, 2, 3, 4, 5, 7, and 8 (Supplementary Table S3). Sequences were aligned using MAFFT [42] and edited using AliView [43]. Those datasets were assessed for the presence of phylogenetic signal by applying the likelihood mapping analysis implemented in the IQ-TREE 1.6.8 software [44]. A maximum likelihood phylogeny was reconstructed using IQ-TREE 1.6.8 software under the HKY+G4 substitution model [44]. We inferred time-scaled trees by using TreeTime [45].

4.7. WNV Epidemiological Data

Human reported cases presenting neurological disease compatible with WNV infection collected between November 2015 and early 2020 were obtained from SINAN. We reinforce the nature of the reports as suspected (not confirmed), being officially defined as cases presenting neurological syndromes compatible with WNV infection, registered as suspected occurrences of West Nile virus infection (code A923). As such, the spatio-temporal series of suspected cases should only be interpreted as a proxy for the possible spatio-temporal dynamics of WNV infections [46].

4.8. Modeling Transmission Potential

To estimate the transmission potential of WNV, we employed the computational approach from Lourenço et al. recently applied in Israel [25] and Portugal [26]. This approach estimates the suitability index P using climatic variables only. The index measures the transmission potential of single adult female mosquitoes (spp. *Culex*) in the animal reservoir and is thus interpreted as a summary measure of the risk for spill-over into human populations. The theory and practice of estimating the index P for mosquito-borne viruses has been previously described in full by Obolski et al. [27]. The epidemiological priors used were the same as in the original study by Lourenço et al. in Israel, which relate to spp. *Culex*, WNV, and an average bird species. Climatic data were obtained from Copernicus.eu (<https://www.copernicus.eu> (accessed on 31 December 2019)); in particular, we used the dataset “essential climate variables for assessment of climate variability from 1979 to present” [47]. This dataset offers climatic variables at a time resolution of 1 month and gridded spatial resolution of 0.25×0.25 .

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/pathogens10070896/s1>, Figure S1: WNV typing tool, Figure S2: Maximum likelihood phylogenetic tree of 2321 WNV complete genomes. Colors indicates different lineages. Highlighted red clade include the WNV viral strain obtained in this study, Table S1: Primer scheme, Table S2: WNV suspected cases reported between 2014–2020 in each Brazilian state, according to SINAN, Table S3: Globally reference WNV sequences from the subset $n = 29$ used in this study.

Author Contributions: Conception and design: É.A.C., M.G., J.L. and L.C.J.A.; Data collection: É.A.C., M.G., L.S.C., V.F., M.A.d.C.e.S.V., D.F.H., D.B.d.A.M., K.R.L.J.C., N.F.O.d.M., A.P.M.R. and L.C.J.A.; Investigations: F.F.A., F.L.L.C., A.M.B.d.F., R.V.d.C., É.A.C., M.G., J.X., V.F., M.I.M.C.G., B.S.Á.d.S.S., A.S.G.S., R.d.P.A.M., N.R.d.C.F., R.F.d.S., R.B.C.T. and J.L.; Data Analysis: M.G., V.F., F.F.A. and J.L.; Writing—Original: É.A.C., M.G., L.S.C., V.F., M.A.d.C.e.S.V., J.L. and L.C.J.A.; Draft Preparation: É.A.C., M.G., L.S.C., V.F., M.A.d.C.e.S.V., J.L. and L.C.J.A.; Revision: É.A.C., M.G., L.S.C., V.F., F.F.A., F.C.d.M.I., M.A.d.C.e.S.V., D.F.H., D.B.d.A.M., M.I.M.C.G., B.S.Á.d.S.S., A.S.G.S., T.d.O., K.R.L.J.C., N.F.O.d.M., A.P.M.R., C.F.C.d.A., L.C.S.F., J.J.M.B., R.B.C.T., O.L.L., S.d.C.S., R.d.P.A.M., R.F.d.S., J.L. and L.C.J.A. Methodology: F.F.A., F.L.L.C. Writing—review & editing: F.F.A., F.L.L.C., A.M.B.d.F., R.V.d.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was founded by CNPq (440685/2016-8, 421598/2018-2), by CAPES (88887.130716/2016-00), by the Pan American Health Organization (IOC-007-FEX-19-2-2-30), by the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), grant number E-26/2002.930/2016 by the European Union’s Horizon 2020 Research and Innovation Programme under ZIKAlliance Grant Agreement no. 734548, by the Horizon 2020 through ZikaPlan and ZikAction (grant agreement numbers 734584 and 734857) and by the National Institutes of Health USA grant U01 AI151698 for the United World Antiviral Research Network (UWARN). MG and LCJA is

supported by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). JL is supported by a lectureship from the Department of Zoology, University of Oxford.

Institutional Review Board Statement: This project was reviewed and approved by the Comissão Nacional de Ética em Pesquisa (CONEP) [National Research Ethics Committee] from the Brazilian Ministry of Health (BrMoH), as part of the arboviral genomic surveillance efforts within the terms of Resolution 510/2016 of CONEP, by the Pan American Health Organization Ethics Review Committee (PAHOERC) (Ref. No. PAHO-2016-08-0029), by the Animal Welfare Committee of Universidade Federal do Piauí, under n°065/19 and by the Oswaldo Cruz Foundation Ethics Committee (CAAE: 90249218.6.1001.5248). All experiments were performed in accordance with relevant guidelines and regulations.

Informed Consent Statement: Not applicable.

Data Availability Statement: Newly generated WNV sequences have been deposited in GenBank under accession numbers MW420987, MW420988 and MW420989.

Acknowledgments: The authors thank the important contributions of the Municipal and Piauí State Health Department (SESAPI, FMS), Municipal and Piauí State Animal Health Department (ADAPI), Laboratório de Saúde Pública do Piauí (LACEN-PI), and the colleague Thiago dos Santos Silva. We also thank the sponsoring institutions: Saint Louis Zoo WildCare Institute and Institute for Conservation Medicine (USA), Universidade Federal do Piauí (UFPI), Fundação de Amparo a Pesquisa do Estado do Piauí (FAPEPI). The authors also thank the Municipal and State Health Department of São Paulo and Minas Gerais state.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fall, G.; Di Paola, N.; Faye, M.; Dia, M.; de Melo Freire, C.C.; Loucoubar, C.; de Andrade Zanotto, P.M.; Faye, O. Biological and phylogenetic characteristics of West African lineages of West Nile virus (DWC Beasley, Ed.). *PLoS Negl. Trop. Dis.* **2017**, *11*, 1–23. [[CrossRef](#)] [[PubMed](#)]
2. Smithburn, K.C.; Hughes, T.P.; Burke, A.W.; Paul, J.H. A neurotropic virus isolated from the blood of a native of Uganda. *Am. J. Trop. Med. Hyg.* **1940**, *20*, 471–472. [[CrossRef](#)]
3. Murgue, B.; Zeller, H.; Deubel, V. The ecology and epidemiology of West Nile virus in Africa, Europe and Asia. *Curr. Top. Microbiol. Immunol.* **2002**, *267*, 195–221. [[PubMed](#)]
4. Campbell, G.L.; Marfin, A.A.; Lanciotti, R.S.; Gubler, D.J. West Nile virus. *Lancet Infect. Dis.* **2002**, *2*, 519–529. [[CrossRef](#)]
5. Gamino, V.; Höfle, U. Pathology and tissue tropism of natural West Nile virus infection in birds: A review. *Vet. Res.* **2013**, *44*, 46–89. [[CrossRef](#)]
6. Bunning, M.L.; Bowen, R.A.; Cropp, B.C.; Sullivan, K.G.; Davis, B.S.; Komar, N.; Godsey, M.; Baker, D.; Hettler, D.L.; Holmes, D.A.; et al. Experimental infection of horses with West Nile virus. *Emerg. Infect. Dis.* **2002**, *8*, 380–386. [[CrossRef](#)]
7. Hayes, E.B.; Sejvar, J.J.; Zaki, S.R.; Lanciotti, R.S.; Bode, A.V.; Campbell, G.L. Virology, pathology, and clinical manifestations of West Nile virus disease. *Emerg. Infect. Dis.* **2005**, *11*, 1174–1179. [[CrossRef](#)]
8. Kramer, L.D.; Li, J.; Shi, P.Y. West Nile virus. *Lancet Neurol.* **2007**, *6*, 171–181. [[CrossRef](#)]
9. Morales, M.A.; Barrandeguy, M.; Fabbri, C.; Garcia, J.B.; Vissani, A.; Trono, K.; Gutierrez, G.; Pigretti, S.; Menchaca, H.; Garrido, N.; et al. West Nile virus isolation from equines in Argentina. *Emerg. Infect. Dis.* **2006**, *12*, 1559–1561. [[CrossRef](#)] [[PubMed](#)]
10. Osorio, J.E.; Ciuoderis, K.A.; Lopera, J.G.; Piedrahita, L.D.; Murphy, D.; LeVasseur, J.; Carrillo, L.; Ocampo, M.C.; Hofmeister, E. Characterization of West Nile viruses isolated from captive American flamingoes (*Phoenicopterus ruber*) in Medellin, Colombia. *Am. J. Trop. Med. Hyg.* **2012**, *87*, 565–572. [[CrossRef](#)]
11. Martins, L.C.; Silva, E.V.; Casseb, L.M.; Silva, S.P.; Cruz, A.C.; Pantoja, J.A.; Medeiros, D.B.; Martins Filho, A.J.; Cruz, E.D.; Araújo, M.T.; et al. First isolation of West Nile virus in Brazil. *Mem. Inst. Oswaldo Cruz* **2019**, *17*, 114–180332. [[CrossRef](#)]
12. Pauvolid-Corrêa, A.; Morales, M.A.; Levis, S.; Figueiredo, L.T.; Couto-Lima, D.; Campos, Z.; Nogueira, M.F.; Silva, E.E.; Nogueira, R.M.; Schatzmayr, H.G. Neutralising antibodies for West Nile virus in horses from Brazilian Pantanal. *Mem. Inst. Oswaldo Cruz* **2011**, *106*, 467–474. [[CrossRef](#)]
13. Vieira, M.A.; Romano, A.P.; Borba, A.S.; Silva, E.V.; Chiang, J.O.; Eulálio, K.D.; Azevedo, R.S.; Rodrigues, S.G.; Almeida-Neto, W.S.; Vasconcelos, P.F. West Nile Virus Encephalitis: The First Human Case Recorded in Brazil. *Am. J. Trop. Med. Hyg.* **2015**, *93*, 377–379. [[CrossRef](#)]
14. Pauvolid-Corrêa, A.; Campos, Z.; Juliano, R.; Velez, J.; Nogueira, R.M.; Komar, N. Serological evidence of widespread circulation of West Nile virus and other flaviviruses in equines of the Pantanal, Brazil. *PLoS Negl. Trop. Dis.* **2014**, *8*, e2706. [[CrossRef](#)] [[PubMed](#)]
15. Morel, A.P.; Webster, A.; Zitelli, L.C.; Umeno, K.; Souza, U.A.; Prusch, F.; Anicet, M.; Marsicano, G.; Bandarra, P.; Trainini, G.; et al. Serosurvey of West Nile virus (WNV) in free-ranging raptors from Brazil. *Braz. J. Microbiol.* **2021**, *52*, 411–418. [[CrossRef](#)]

16. Melandri, V.; Guimarães, A.É.; Komar, N.; Nogueira, M.L.; Mondini, A.; Fernandez-Sesma, A.; Alencar, J.; Bosch, I. Serological detection of West Nile virus in horses and chicken from Pantanal, Brazil. *Mem. Inst. Oswaldo Cruz* **2012**, *107*, 1073–1075. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Ometto, T.; Durigon, E.L.; de Araujo, J.; Aprelon, R.; de Aguiar, D.M.; Cavalcante, G.T.; Melo, R.M.; Levi, J.E.; de Azevedo Júnior, S.M.; Petry, M.V.; et al. West Nile virus surveillance, Brazil, 2008–2010. *Trans. R. Soc. Trop. Med. Hyg.* **2013**, *107*, 723–730. [\[CrossRef\]](#)
18. Shocket, M.S.; Verwillow, A.B.; Numazu, M.G.; Slamani, H.; Cohen, J.M.; El Moustaid, F.; Rohr, J.; Johnson, L.R.; Mordecai, E.A. Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures between 23 °C and 26 °C. *Elife* **2020**, *9*, e58511. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Haussig, J.M.; Young, J.J.; Gossner, C.M.; Mezei, E.; Bella, A.; Sirbu, A.; Pervanidou, D.; Drakulovic, M.B.; Sudre, B. Early start of the West Nile fever transmission season 2018 in Europe. *Eurosurveillance* **2018**, *23*. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Riccardo, F.; Bolici, F.; Fafangel, M.; Jovanovic, V.; Socan, M.; Klepac, P.; Plavska, D.; Vasic, M.; Bella, A.; Diana, G.; et al. West Nile virus in Europe: After action reviews of preparedness and response to the 2018 transmission season in Italy, Slovenia, Serbia and Greece. *Global Health* **2020**, *16*, 47. [\[CrossRef\]](#)
21. Bakonyi, T.; Haussig, J.M. West Nile virus keeps on moving up in Europe. *Eurosurveillance* **2020**, *25*. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Pfeffer, M.; Proebster, B.; Kinney, R.M.; Kaaden, O.R. Genus-specific detection of alphaviruses by a semi-nested reverse transcription-polymerase chain reaction. *Am. J. Trop. Med. Hyg.* **1997**, *57*, 709–718. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Castro-Jorge, L.A.; Siconelli, M.J.L.; Ribeiro, B.D.S.; Moraes, F.M.; Moraes, J.B.; Agostinho, M.R.; Klein, T.M.; Floriano, V.G.; Fonseca, B.A.L.D. West Nile virus infections are here! Are we prepared to face another flavivirus epidemic? *Rev. Soc. Bras. Med. Trop.* **2019**, *52*, e20190089. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Hernandez-Triana, L.M.; Jeffries, C.L.; Mansfield, K.L.; Carnell, G.; Fooks, A.R.; Johnson, N. Emergence of West Nile virus lineage 2 in Europe: A review on the introduction and spread of a mosquito-borne disease. *Front. Public Health* **2014**, *2*, 271. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Lourenço, J.; Thompson, R.N.; Thézé, J.; Obolski, U. Characterising West Nile virus epidemiology in Israel using a transmission suitability index. *Eurosurveillance* **2020**, *2*, 5–41.
26. Lourenco, J.; Barros, S.C.; Ze-Ze, L.; Damineli, D.S.; Giovanetti, M.; Osorio, H.C.; Amaro, F.; Henriques, A.M.; Ramos, F.; Luis, T.; et al. West Nile virus in Portugal. *MedRxiv* **2021**. [\[CrossRef\]](#)
27. Obolski, U.; Perez, P.N.; Villabona-Arenas, C.J.; Thézé, J.; Faria, N.R.; Lourenço, J. MVSE: An R-package that estimates a climate-driven mosquito-borne viral suitability index. *Methods Ecol. Evol.* **2019**, *10*, 1357–1370. [\[CrossRef\]](#)
28. Hadfield, J.; Brito, A.F.; Swetnam, D.M.; Vogels, C.B.F.; Tokarz, R.E.; Andersen, K.G.; Smith, R.C.; Bedford, T.; Grubaugh, N.D. Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. *PLoS Pathog.* **2019**, *15*, e1008042. [\[CrossRef\]](#)
29. Viana, D.S.; Santamariá, L.; Figuerola, J. Migratory birds as global dispersal vectors. *Trends Ecol. Evol.* **2016**, *31*, 763–775. [\[CrossRef\]](#)
30. Siconelli, M.J.L.; Jorge, D.M.M.; Castro-Jorge, L.A.; Fonseca-Júnior, A.A.; Nascimento, M.L.; Floriano, V.G.; Souza, F.R.; Queiroz-Júnior, E.M.; Camargos, M.F.; Costa, E.D.L.; et al. Evidence for current circulation of an ancient West Nile virus strain (NY99) in Brazil. *Rev. Soc. Bras. Med. Trop.* **2021**, *54*, e0687–e2020. [\[CrossRef\]](#)
31. Petrone, M.E.; Earnest, R.; Lourenço, J.; Kraemer, M.U.G.; Paulino-Ramirez, R.; Grubaugh, N.D.; Tapia, L. Asynchronicity of endemic and emerging mosquito-borne disease outbreaks in the Dominican Republic. *Nat. Commun.* **2021**, *12*, 151. [\[CrossRef\]](#)
32. Fulop, L.; Barrett, A.D.; Philippotts, R.; Martin, K.; Leslie, D.; Titball, R.W. Rapid identification of flaviviruses based on conserved NS5 gene sequences. *J. Virol. Methods* **1993**, *44*, 179–188. [\[CrossRef\]](#)
33. Silva, A.S.G.; Matos, A.C.D.; da Cunha, M.A.C.R.; Rehfeld, I.S.; Galinari, G.C.F.; Marcelino, S.A.C.; Saraiva, L.H.G.; Martins, N.R.D.S.; Maranhão, R.P.A.; Lobato, Z.I.P.; et al. West Nile virus associated with equid encephalitis in Brazil, 2018. *Transbound Emerg Dis.* **2019**, *66*, 445–453. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Costa, E.A.; Rosa, R.; Oliveira, T.S.; Assis, A.C.; Paixão, T.A.; Santos, R.L. Molecular characterization of neuropathogenic Equine Herpesvirus 1 Brazilian isolates. *Arq. Bras. Med. Vet. Zootec.* **2015**, *67*, 1183–1187. [\[CrossRef\]](#)
35. Sorg, I.; Metzler, A. Detection of Borna Disease Virus RNA in Formalin-Fixed, Paraffin-Embedded Brain Tissues by Nested PCR. *J. Clin. Microbiol.* **1995**, *4*, 821–823. [\[CrossRef\]](#)
36. Faria, N.R.; Quick, J.; Claro, I.M.; Theze, J.; de Jesus, J.G.; Giovanetti, M.; Kraemer, M.U.; Hill, S.C.; Black, A.; da Costa, A.C.; et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **2017**, *546*, 406–410. [\[CrossRef\]](#)
37. Quick, J.; Grubaugh, N.D.; Pullan, S.T.; Claro, I.M.; Smith, A.D.; Gangavarapu, K.; Oliveira, G.; Robles-Sikisaka, R.; Rogers, T.F.; Beutler, N.A.; et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **2017**, *12*, 1261. [\[CrossRef\]](#)
38. Vilsker, M.; Moosa, Y.; Nooij, S.; Fonseca, V.; Ghysens, Y.; Dumon, K.; Pauwels, R.; Alcantara, L.C.; Vanden Eynden, E.; Vandamme, A.M.; et al. Genome Detective: An automated system for virus identification from high-throughput sequencing data. *Bioinformatics* **2019**, *35*, 871–873. [\[CrossRef\]](#)
39. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *21*, 403–410. [\[CrossRef\]](#)

40. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [[CrossRef](#)] [[PubMed](#)]
41. Lemey, P.; Salemi, M.; Vandamme, A. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
42. Katoh, K.; Kuma, K.I.; Toh, H.; Miyata, T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **2005**, *33*, 511–518. [[CrossRef](#)] [[PubMed](#)]
43. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics* **2014**, *30*, 3276–3278. [[CrossRef](#)] [[PubMed](#)]
44. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [[CrossRef](#)] [[PubMed](#)]
45. Sagulenko, P.; Puller, V.; Neher, R.A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **2018**, *4*, vex042. [[CrossRef](#)] [[PubMed](#)]
46. Ministério da Saúde (MS). Secretaria de Vigilância em Saúde. In *Monitoramento da Febre do Nilo Ocidental no Brasil, 2014 a 2019 (Nota Informativa)*; MS: Brasília, Brazil, 2019; 7p. Available online: <https://antigo.saude.gov.br/images/pdf/2019/julho/08/informe-febre-niloocidental-n1-8jul19b.pdf> (accessed on 31 December 2019).
47. Copernicus Climate Data Store. Available online: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecv-for-climate-change?tab=overview> (accessed on 23 December 2020).

CHAPTER 6: SYNTHESIS OF RESEARCH FINDINGS

7.1 Key Themes

The application of computational tools contributed to a better understanding of the infections caused by emerging and re-emerging viruses, thus contributing to the resolution of issues relevant to public health.

In the group of emerging and re-emerging infectious diseases in Brazil, infections caused by arboviruses DENV, ZIKV, CHIKV and YFV pose a major threat. It is therefore crucial to control the epidemics arising from transmission of these viruses. The occurrence of a recent ZIKV epidemic associated with cases of microcephaly and congenital infections, the emergence of CHIKV associated with its limiting capacity for long periods, and the re-emergence of YFV in the wild cycle, generating an epidemic beyond the limits of the Amazon, all highlight the need for systematic and continuous monitoring of these arboviruses.

The generated genomic data obtained in those studies provided a more detailed understanding of the introduction and progression of several arboviruses currently circulating and co-circulating within Brazilian regions revealing the timing, source, and likely routes of their transmission and dispersion. These findings helped guide future efforts on viral surveillance and Health authorities were able to use our inferences to strengthen vaccination programs and prioritize surveillance and control strategies. The results obtained have been translated into modelling efforts on other enzootic viruses, thus supporting public health decisions to prevent and cope with future outbreaks. Finally, the results of this thesis contributed to the establishment of an International Network for genomic monitoring of neglected, emerging and re-emerging viral pathogen, playing an important role in solving pressing public health issues. Studies involving more in-depth molecular analyses of circulating strains can help the Health Authorities to adopt appropriate measures to control epidemics, and to monitor the dynamics and spread of novel viruses.

Studies which involve more in-depth molecular analysis of sequences circulating in the country and which track viral dispersion/circulation could guide the use of appropriate measures to control the epidemic and monitor the dynamics of the evolution of circulating viral strains. In addition, the genomic data generated in this study shed light on the potential impact of additional mutations, providing support for appropriate preventive actions and therapeutic interventions.

Implementing a standard protocol using the MinION portable sequencer to identify potential emerging viruses has become of fundamental importance in order to allow rapid identification and monitoring of possible introductions into different regions. In addition, it has enabled us to investigate

the evolution of viral genomes, better understand the origin of outbreaks and epidemics, and maintain diagnostic methods.

The bioinformatics tools we developed, namely “GenomeDetective MinION module, Dengue Typingtool, Chikungunya Typingtool and Zika Typingtool” have enabled viruses obtained through the rapid and accurate assembly, identification, and classification of viruses present in samples obtained through state-of-the-art sequencing. The tools provide an aesthetic, graphic user-friendly interface and produce comprehensive reports that serve various purposes, such as presenting information on viral species, visual representation of genomic components, and visualization of alignments. They are able to measure nucleotide and amino acid similarity, and to determine read counts in a BAM file using an implementation of the Burrows-Wheeler Aligner software (BWA) [78]. Furthermore, the tools can be accessed directly from other viral assembly and identification tools through an application programming interface (API), as for example, from Genome Detective [79] a short-read sequencing assembly pipeline. In this way, users can quickly obtain reliable and useful information from the raw viral sequencing data through the MinION portable platform. The FastQ data obtained from the Nanopore sequencing can be submitted to the Genome detective pipeline after running the basecalling and demultiplexing processes, which first convert the FAST5 data into FastQ and subsequently translate raw signals (referred to as squiggle) into nucleotide sequences. It is not necessary to install any additional software or possess in-depth knowledge of bioinformatics tools. It is also possible to obtain visual confirmation of the identified viruses, with high specificity and sensitivity, based on curated reference databases.

The genotyping tools that we developed in Chapter 3 to 5 are able to perform genomic characterization for DENV, ZIKV, CHIKV, WNV, Coronavirus and SARS-CoV-2 quickly and efficiently. This method involves aligning a query string 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 string is assigned the genotype of the reference strain with the highest bootstrap scores ($>70\%$). This provides the user with information about the genotype that is circulating in a given region from sequenced samples. The tool also provides phylogenetic trees with groupings of monophyletic clades and is able to analyse up to 2000 thousand sequences at a time ranging from sequence fragments to complete genomes.

The tools developed within the scope of this thesis, which appear in Chapters 2 to 5, contributed to the generation of a total of 291 complete arbovirus genomes. These included 49 CHIKV, 61 DENV-1, 170 DENV-2 and 11 DENV-4 genomes all through MinION sequencing and more than 15 thousand genome SARS-CoV-2 between Illimuna and MinION [80–86]. These sequences significantly increased the number of complete sequences deposited in public databases and allowed us to make phylogenetic inferences about the dates of introduction of the lineages circulating in different places.

For example, findings from the show that the ECSA genotype in the state of Rio de Janeiro accounted for 66.8% and 69.7% of cases accumulated in the Southeast region of Brazil in 2016 and 2018, respectively [87]. In Mato Grosso, Brazil, the findings of the phylogenetic analysis revealed a complex pattern of CHIKV transmission between epidemic seasons and sampled locations. This suggests that Brazil has played a critical role in seeding international dispersions of arboviral infections to other countries in the Americas, such as Paraguay and Haiti [80,88]. From DENV sequences generated in Brazil and Paraguay, the phylogenetic analyses show that the Southeast and North regions of Brazil and Paraguay, which are all major tourist destinations in South America, were fundamental in the dispersion of DENV-1 and DENV-2 [82,88]. However, the Caribbean also played an important role in the spread of the virus across Paraguay. The results show that intelligent use of bioinformatics tools can reveal intricate transmission dynamics between sampled locations, especially when combined with epidemiological data and travel history.

Therefore, the urban cycles of DENV and CHIKV have been responsible for large epidemics in continents that cover several countries [89,90]. After causing outbreaks on the Indian Ocean Islands and in India, the viruses also spread to Asia [89,90]. DENV has been endemic in the Americas since the 1980s, with a seasonal cycle causing more than 20 million cases there [91,92]. In 2013, the introduction of CHIKV in the Caribbean resulted in an epidemic in St. Martin, followed by its spread to other islands nearby [93,94]. At the same time, CHIKV also spread throughout South and Central America, causing significant epidemics throughout the region where competent vectors had already been established. Studies on vector competence demonstrate that this can be highly variable in natural populations and is determined by genotype-genotype interactions, in which successful transmission depends on a specific combination of genetic characteristics of mosquitoes and viruses under specific environmental conditions. Factors such as conducive temperatures and daily fluctuations play a fundamental role in the competence of vectors in relation to the multiplication of pathogens [90,95].

The subtyping tools in Chapters 3 to 5 classified the genomes generated belong to the ECSA genotype of CHIKV, DENV-1 genotype V, DENV-2 genotype III, and DENV-4 genotype II [82], which was expected, as previously described [96–100]. Although CHIKV is related to explosive outbreaks worldwide [101], we report the persistence of ECSA genotype circulation within the South American population. Thus, we conclude that the establishment of a network of laboratories working together to provide rapid responses to the emergence of epidemics is essential for its management.

7.2 Recommendations for policy

This study showed that it is possible to carry out research and promote capacity building of Central Public Health Laboratories (LACENS - Laboratórios Centrais de Saúde Pública) in Brazil through the exchange of experiences and knowledge on modern DNA sequencing technologies and

genomic data analysis. Thus, since its conception, the proposal presented here has included a broad educational component.

This study described the results of a training course in next genome sequencing (NGS) techniques and bioinformatics analysis designed for professionals from public health laboratories in Brazil. Such activities with participants from outside of academia can improve health workers and managers' qualifications in sequencing technologies and genomic data analysis of circulating and co-circulating arboviruses.

Training like this can help staff to identify circulating viral strains, characterize infected individuals, produce new diagnostic techniques and immunotherapeutic approaches. These include seeking to better understand the transmission chains of these arboviruses and identify the origin and spread of these infections.

It is important that health authorities be given the training they need to make informed decisions so they can respond appropriately to outbreaks and epidemics. Laboratories like LACENS can assist in genomic surveillances of arboviruses and other public health diseases like influenza and measles. In this way, they help generate data for the composition and update of the vaccine carried out annually by the Ministry of Health and the World Health Organization for the influenza vaccine.

In addition, as demonstrated in this research, laboratories can also play a vital role in capacity building and the training health professionals, which substantially contribute to more effective and assertive responses by health surveillance.

7.3 Publications Declaration

7.3.1 Papers

1. Giovanetti M, **Fonseca V**, Wilkinson E, Tegally H, San EJ, Althaus CL, A, et al. Replacement of the Gamma by the Delta variant in Brazil: impact of lineage displacement on the ongoing pandemic, *Virus Evolution*, 2022, veac024, <https://doi.org/10.1093/ve/veac024>.
2. Giovanetti M, Pereira LA, Adelino TÉR, **Fonseca V**, Xavier J, de Araújo Fabri A, et al. A retrospective overview of zika virus evolution in the midwest of brazil. *Microbiol Spectr*. 2022 Mar 7;e0015522.
3. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, ..., **Fonseca V**, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science*. 2021 Oct 22;374(6566):423–31.
4. Kashima S, Slavov SN, Giovanetti M, Rodrigues ES, Patané JSL, Viala VL, ..., **Fonseca V**, et al. Introduction of SARS-CoV-2 C.37 (WHO VOI lambda) in the Sao Paulo State, Southeast Brazil. *J Med Virol*. 2021 Oct 14;
5. de Lima STS, de Souza WM, Cavalcante JW, da Silva Candido D, Fumagalli MJ, Carrera J-P, ..., **Fonseca V**, et al. Fatal outcome of chikungunya virus infection in brazil. *Clin Infect Dis*. 2021 Oct 5;73(7):e2436–43.
6. **Fonseca V**, de Jesus R, Adelino T, Reis AB, de Souza BB, Ribeiro AA, et al. Genomic evidence of SARS-CoV-2 reinfection case with the emerging B.1.2 variant in Brazil. *J Infect*. 2021 Aug;83(2):237–79.
7. Tosta S, Giovanetti M, Brandão Nardy V, Reboredo de Oliveira da Silva L, Kelly Astete Gómez

- M, Gomes Lima J, ..., **Fonseca V**, et al. Short Report: Early genomic detection of SARS-CoV-2 P.1 variant in Northeast Brazil. *PLoS Negl Trop Dis*. 2021 Jul 19;15(7):e0009591.
8. Costa ÉA, Giovanetti M, Silva Catenacci L, **Fonseca V**, Aburjaile FF, Chalhoub FLL, et al. West Nile virus in Brazil. *Pathogens*. 2021 Jul 15;10(7).
 9. Slavov SN, Patané JSL, Bezerra RDS, Giovanetti M, **Fonseca V**, Martins AJ, et al. Genomic monitoring unveils the early detection of the SARS-CoV-2 B.1.351 (beta) variant (20H/501Y.V2) in Brazil. *J Med Virol*. 2021 Jul 9;
 10. Angeletti S, Giovanetti M, Fogolari M, De Florio L, Francesconi M, Veralli R, ..., **Fonseca V**, et al. Detection of a SARS-CoV-2 P.1.1 variant lacking N501Y in a vaccinated health care worker in Italy. *J Infect*. 2021 Jul 6;
 11. Iani FCM, Giovanetti M, **Fonseca V**, Souza WM, Adelino TER, Xavier J, et al. Epidemiology and evolution of Zika virus in Minas Gerais, Southeast Brazil. *Infect Genet Evol*. 2021 Jul;91:104785.
 12. Giovanetti M, Alcantara LCJ, Dorea AS, Ferreira QR, Marques W de A, Junior Franca de Barros J, ..., **Fonseca V**, et al. Promoting responsible research and innovation (RRI) during Brazilian activities of genomic and epidemiological surveillance of arboviruses. *Front Public Health*. 2021 Jul 1;9:693743.
 13. Slavov SN, Giovanetti M, Dos Santos Bezerra R, **Fonseca V**, Santos EV, Rodrigues ES, et al. Molecular surveillance of the on-going SARS-CoV-2 epidemic in Ribeirão Preto City, Brazil. *Infect Genet Evol*. 2021 Jun 24;93:104976.
 14. Pereira F, Tosta S, Lima MM, Reboredo de Oliveira da Silva L, Nardy VB, Gómez MKA, ..., **Fonseca V**, et al. Genomic surveillance activities unveil the introduction of the SARS-CoV-2 B.1.525 variant of interest in Brazil: Case report. *J Med Virol*. 2021 May 15;
 15. Lopes EN, **Fonseca V**, Frias D, Tosta S, Salgado Á, Assunção Vialle R, et al. Betacoronavirus genome analysis reveals evolution toward specific codon usage: Implications for SARS-CoV-2 mitigation strategies. *J Med Virol*. 2021 May 2;
 16. Gräf T, Vazquez C, Giovanetti M, de Bruycker-Nogueira F, **Fonseca V**, Claro IM, et al. Epidemiologic history and genetic diversity origins of chikungunya and dengue viruses, Paraguay. *Emerging Infect Dis*. 2021 May;27(5):1393–404.
 17. Nonaka CKV, Franco MM, Gräf T, de Lorenzo Barcia CA, de Ávila Mendonça RN, de Sousa KAF, ..., **Fonseca V**, et al. Genomic Evidence of SARS-CoV-2 Reinfection Involving E484K Spike Mutation, Brazil. *Emerging Infect Dis*. 2021 May;27(5):1522–4.
 18. Giovanetti M, Cella E, Benedetti F, Rife Magalis B, **Fonseca V**, Fabris S, et al. SARS-CoV-2 shifting transmission dynamics and hidden reservoirs potentially limit efficacy of public health interventions in Italy. *Commun Biol*. 2021 Apr 21;4(1):489.
 19. de Oliveira EC, **Fonseca V**, Xavier J, Adelino T, Morales Claro I, Fabri A, et al. Short report: Introduction of chikungunya virus ECSA genotype into the Brazilian Midwest and its dispersion through the Americas. *PLoS Negl Trop Dis*. 2021 Apr 16;15(4):e0009290.
 20. Adelino TÉR, Giovanetti M, **Fonseca V**, Xavier J, de Abreu AS, do Nascimento VA, et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. *Nat Commun*. 2021 Apr 16;12(1):2296.
 21. Xavier J, **Fonseca V**, Bezerra JF, do Monte Alves M, Mares-Guia MA, Claro IM, et al. Chikungunya virus ECSA lineage reintroduction in the northeasternmost region of Brazil. *Int J Infect Dis*. 2021 Apr;105:120–3.
 22. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, ..., **Fonseca V**, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat Med*. 2021 Mar;27(3):440–6.
 23. Torres MC, Lima de Mendonça MC, Damasceno Dos Santos Rodrigues C, **Fonseca V**, Ribeiro MS, Brandão AP, et al. Dengue Virus Serotype 2 Intrahost Diversity in Patients with Different Clinical Outcomes. *Viruses*. 2021 Feb 23;13(2).
 24. Giandhari J, Pillay S, Wilkinson E, Tegally H, Sinayskiy I, Schuld M, ..., **Fonseca V**, et al. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int J Infect Dis*. 2021 Feb;103:234–41.
 25. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, **Fonseca V**, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. 2021;592(7854):438–43.

26. San JE, Ngcapu S, Kanzi AM, Tegally H, **Fonseca V**, Giandhari J, et al. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol.* 2021 Jan;7(1):veab041.
27. Xavier J, Giovanetti M, Adelino T, **Fonseca V**, Barbosa da Costa AV, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg Microbes Infect.* 2020 Dec;9(1):1824–34.
28. Fabri AA, Rodrigues CDDS, Santos CCD, Chalhoub FLL, Sampaio SA, Faria NR da C, ..., **Fonseca V**, et al. Co-Circulation of Two Independent Clades and Persistence of CHIKV-ECSA Genotype during Epidemic Waves in Rio de Janeiro, Southeast Brazil. *Pathogens.* 2020 Nov 26;9(12).
29. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, ..., **Fonseca V**, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel).* 2020 Aug 17;11(8).
30. Goes de Jesus J, Gräf T, Giovanetti M, Mares-Guia MA, Xavier J, Lima Maia M, ..., **Fonseca V**, et al. Yellow fever transmission in non-human primates, Bahia, Northeastern Brazil. *PLoS Negl Trop Dis.* 2020 Aug 11;14(8):e0008405.
31. Cleemput S, Dumon W, **Fonseca V**, Abdool Karim W, Giovanetti M, Alcantara LC, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics.* 2020 Jun 1;36(11):3552–5.
32. Giovanetti M, Faria NR, Lourenço J, Goes de Jesus J, Xavier J, Claro IM, ..., **Fonseca V**, et al. Genomic and epidemiological surveillance of zika virus in the amazon region. *Cell Rep.* 2020 Feb 18;30(7):2275–2283.e7.
33. Morais-Rodrigues F, Silv Erio-Machado R, Kato RB, Rodrigues DLN, Valdez-Baez J, **Fonseca V**, et al. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene.* 2020 Feb 5;726:144168.
34. Goes de Jesus J, da Luz Wallau G, Lima Maia M, Xavier J, Oliveira Lima MA, **Fonseca V**, et al. Persistence of chikungunya ECSA genotype and local outbreak in an upper medium class neighborhood in Northeast Brazil. *PLoS One.* 2020 Jan 8;15(1):e0226098.
35. Pereira Gusmão Maia Z, Mota Pereira F, do Carmo Said RF, **Fonseca V**, Gräf T, de Bruycker Nogueira F, et al. Return of the founder Chikungunya virus to its place of introduction into Brazil is revealed by genomic characterization of exanthematic disease cases. *Emerg Microbes Infect.* 2020;9(1):53–7.
36. Giovanetti M, de Mendonça MCL, **Fonseca V**, Mares-Guia MA, Fabri A, Xavier J, et al. Yellow Fever Virus Reemergence and Spread in Southeast Brazil, 2016-2019. *J Virol.* 2019 Dec 12;94(1).
37. Xavier J, Giovanetti M, **Fonseca V**, Thézé J, Gräf T, Fabri A, et al. Circulation of chikungunya virus East/Central/South African lineage in Rio de Janeiro, Brazil. *PLoS One.* 2019 Jun 11;14(6):e0217871.
38. **Fonseca V**, Libin PJK, Theys K, Faria NR, Nunes MRT, Restovic MI, et al. A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes. *PLoS Negl Trop Dis.* 2019 May 8;13(5):e0007231.
39. Vilsker M, Moosa Y, Nooij S, **Fonseca V**, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics.* 2019 Mar 1;35(5):871–3.
40. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, **Fonseca V**, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol.* 2019;10:3119.
41. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, ..., **Fonseca V**, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science.* 2018 Aug 31;361(6405):894–9.

7.3.2 Book

1. Depa L, Depa L, Vasconcelos C, **Fonseca V**, Frias D. Estudo do uso de códons nos vírus da Dengue, Zika e Chikungunya com foco em terapia por inibição seletiva de tRNAs contra

arboviroses. Mariano D, editor. Alfahelix; 2021.

7.3.3 Chapter

1. Xavier J, Tosta S, Adelino T, **Fonseca V**, Giovanetti M, Alcantara LCJ. Classification of Zika virus sequences with respect to their species and subspecies. *Zika virus impact, diagnosis, control, and models*. Elsevier; 2021. p. 29–37.
2. Giovanetti M, Salgado A, **Fonseca V**, Tosta F de O, Xavier J, de Jesus JG, et al. Pan-genomics of virus and its applications. *Pan-genomics: Applications, Challenges, and Future Prospects*. Elsevier; 2020. p. 237–50.
3. **Fonseca V**, Xavier J, Emmanuel James S, de Oliveira T, Maria Bispo de Filippis A, Carlos Junior Alcantara L, et al. Mosquito-Borne Viral Diseases: Control and Prevention in the Genomics Era. *Current Topics in the Epidemiology of Vector-Borne Diseases [Working Title]*. IntechOpen; 2020.

7.4 References

1. Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. Epidemic arboviral diseases: priorities for research and public health. *Lancet Infect Dis.* 2017;17: e101–e106. doi:10.1016/S1473-3099(16)30518-7
2. Cardoso CW, Paploski IAD, Kikuti M, Rodrigues MS, Silva MMO, Campos GS, et al. Outbreak of Exanthematous Illness Associated with Zika, Chikungunya, and Dengue Viruses, Salvador, Brazil. *Emerging Infect Dis.* 2015;21: 2274–2276. doi:10.3201/eid2112.151167
3. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature.* 2017;546: 406–410. doi:10.1038/nature22401
4. Mota MT de O, Ribeiro MR, Vedovello D, Nogueira ML. Mayaro virus: a neglected arbovirus of the Americas. *Future Virol.* 2015;10: 1109–1122. doi:10.2217/fvl.15.76
5. Long KC, Ziegler SA, Thangamani S, Hausser NL, Kochel TJ, Higgs S, et al. Experimental transmission of Mayaro virus by *Aedes aegypti*. *Am J Trop Med Hyg.* 2011;85: 750–757. doi:10.4269/ajtmh.2011.11-0359
6. Serra OP, Cardoso BF, Ribeiro ALM, Santos FAL dos, Silhessarenko RD. Mayaro virus and dengue virus 1 and 4 natural infection in culicids from Cuiabá, state of Mato Grosso, Brazil. *Mem Inst Oswaldo Cruz.* 2016;111: 20–29. doi:10.1590/0074-02760150270
7. Wiggins K, Eastmond B, Alto BW. Transmission potential of Mayaro virus in Florida *Aedes aegypti* and *Aedes albopictus* mosquitoes. *Med Vet Entomol.* 2018;32: 436–442. doi:10.1111/mve.12322
8. Moratorio G, Vignuzzi M. Monitoring and redirecting virus evolution. *PLoS Pathog.* 2018;14: e1006979. doi:10.1371/journal.ppat.1006979
9. Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. *Proc Biol Sci.* 2015;282: 20142878. doi:10.1098/rspb.2014.2878
10. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature.* 2016;530: 228–232. doi:10.1038/nature16996
11. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc.* 2017;12: 1261–1276. doi:10.1038/nprot.2017.066
12. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science.* 2018;361: 894–899. doi:10.1126/science.aat7115
13. Brooks GF. *Microbiologia Medica: de Jawetz, Melnick & Adelberg.* Cincias Biol—gicas e Naturais. Mc Graw Hill; 2014.
14. Figueiredo LTM. Emergent arboviruses in Brazil. *Rev Soc Bras Med Trop.* 2007;40: 224–229. doi:10.1590/s0037-86822007000200016
15. Cleton N, Koopmans M, Reimerink J, Godeke G-J, Reusken C. Come fly with me: review of clinically important arboviruses for global travelers. *J Clin Virol.* 2012;55: 191–203. doi:10.1016/j.jcv.2012.07.004
16. Gubler DJ. Human arbovirus infections worldwide. *Ann N Y Acad Sci.* 2001;951: 13–24. doi:10.1111/j.1749-6632.2001.tb02681.x
17. Mukhopadhyay S, Kuhn RJ, Rossmann MG. A structural perspective of the flavivirus life cycle. *Nat Rev Microbiol.* 2005;3: 13–22. doi:10.1038/nrmicro1067
18. Cook S, Holmes EC. A multigene analysis of the phylogenetic relationships among

- the flaviviruses (Family: Flaviviridae) and the evolution of vector transmission. *Arch Virol.* 2006;151: 309–325. doi:10.1007/s00705-005-0626-6
19. Lindenbach BD. The viruses and their replication. In: Knipe, DM and Howley PM, Eds, *Fields Virology*, Lippincott Williams and Wilkins. 5rd ed. 2007. p. 1101.
 20. Cook S, Moureau G, Harbach RE, Mukwaya L, Goodger K, Ssenfuka F, et al. 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.* 2009;90: 2669–2678. doi:10.1099/vir.0.014183-0
 21. LINDENBACH BD, RICE CM. Molecular Biology of the Flaviviruses. In: Chambers T, Monath T, editors. *The Flaviviruses: Structure, Replication and Evolution*. 1st Edition. California, Academic Press; 2003. pp. 23–61.
 22. Hoshino K, Isawa H, Tsuda Y, Sawabe K, Kobayashi M. Isolation and characterization of a new insect flavivirus from *Aedes albopictus* and *Aedes flavopictus* mosquitoes in Japan. *Virology.* 2009;391: 119–129. doi:10.1016/j.virol.2009.06.025
 23. Sang RC, Gichogo A, Gachoya J, Dunster MD, Ofula V, Hunt AR, 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.* 2003;148: 1085–1093. doi:10.1007/s00705-003-0018-8
 24. Pierson TC, Kielian M. Flaviviruses: braking the entering. *Curr Opin Virol.* 2013;3: 3–12. doi:10.1016/j.coviro.2012.12.001
 25. Cook S, Moureau G, Kitchen A, Gould EA, de Lamballerie X, Holmes EC, et al. Molecular evolution of the insect-specific flaviviruses. *J Gen Virol.* 2012;93: 223–234. doi:10.1099/vir.0.036525-0
 26. Belshaw R, de Oliveira T, Markowitz S, Rambaut A. The RNA virus database. *Nucleic Acids Res.* 2009;37: D431-5. doi:10.1093/nar/gkn729
 27. Moureau G, Ninove L, Izri A, Cook S, De Lamballerie X, Charrel RN. Flavivirus RNA in phlebotomine sandflies. *Vector Borne Zoonotic Dis.* 2010;10: 195–197. doi:10.1089/vbz.2008.0216
 28. Chambers TJ, Hahn CS, Galler R, Rice CM. Flavivirus genome organization, expression, and replication. *Annu Rev Microbiol.* 1990;44: 649–688. doi:10.1146/annurev.mi.44.100190.003245
 29. Sánchez-Seco MP, Rosario D, Domingo C, Hernández L, Valdés K, Guzmán MG, et al. Generic RT-nested-PCR for detection of flaviviruses using degenerated primers and internal control followed by sequencing for specific identification. *J Virol Methods.* 2005;126: 101–109. doi:10.1016/j.jviromet.2005.01.025
 30. Harris E, Holden KL, Edgil D, Polacek C, Clyde K. Molecular biology of flaviviruses. *Novartis Found Symp.* 2006;277: 23-39; discussion 40, 71. doi:10.1002/0470058005.ch3
 31. HOLMES E, TWIDDY S. The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol.* 2003;3: 19–28. doi:10.1016/S1567-1348(03)00004-2
 32. Hammon WM, Rudnick A, Sather G, Rogers KD, Morse LJ. New hemorrhagic fevers of children in the Philippines and Thailand. *Trans Assoc Am Physicians.* 1960;73: 140–155.
 33. Mairuhu ATA, Wagenaar J, Brandjes DPM, van Gorp ECM. Dengue: an arthropod-borne disease of global importance. *Eur J Clin Microbiol Infect Dis.* 2004;23: 425–433. doi:10.1007/s10096-004-1145-1
 34. Silva AM da. Molecular characterization of dengue circulating in Pernambuco: epidemiological implications [Internet]. Doctoral dissertation. 2013. Available: <https://www.arca.fiocruz.br/bitstream/icict/10514/1/172.pdf>

35. Rigau-Pérez JG, Clark GG, Gubler DJ, Reiter P, Sanders EJ, Vorndam AV. Dengue and dengue haemorrhagic fever. *Lancet*. 1998;352: 971–977. doi:10.1016/s0140-6736(97)12483-7
36. Gubler DJ. Dengue and dengue hemorrhagic fever. *Clin Microbiol Rev*. 1998;11: 480–496. doi:10.1128/CMR.11.3.480
37. Gubler DJ. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends Microbiol*. 2002;10: 100–103. doi:10.1016/s0966-842x(01)02288-0
38. Weaver SC, Vasilakis N. Molecular evolution of dengue viruses: contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. *Infect Genet Evol*. 2009;9: 523–540. doi:10.1016/j.meegid.2009.02.003
39. Halstead SB. Dengue vaccine development: a 75% solution? *Lancet*. 2012;380: 1535–1536. doi:10.1016/S0140-6736(12)61510-4
40. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res*. 2012;40: D593-8. doi:10.1093/nar/gkr859
41. Dick GWA, Kitchen SF, Haddock AJ. Zika virus. I. Isolations and serological specificity. *Trans R Soc Trop Med Hyg*. 1952;46: 509–520. doi:10.1016/0035-9203(52)90042-4
42. Macnamara FN. Zika virus: a report on three cases of human infection during an epidemic of jaundice in Nigeria. *Trans R Soc Trop Med Hyg*. 1954;48: 139–145. doi:10.1016/0035-9203(54)90006-1
43. Marchette NJ, Garcia R, Rudnick A. Isolation of Zika virus from *Aedes aegypti* mosquitoes in Malaysia. *Am J Trop Med Hyg*. 1969;18: 411–415. doi:10.4269/ajtmh.1969.18.411
44. Olson JG, Ksiazek TG, Suhandiman, Triwibowo. Zika virus, a cause of fever in Central Java, Indonesia. *Trans R Soc Trop Med Hyg*. 1981;75: 389–393. doi:10.1016/0035-9203(81)90100-0
45. Duffy MR, Chen T-H, Hancock WT, Powers AM, Kool JL, Lanciotti RS, et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. *N Engl J Med*. 2009;360: 2536–2543. doi:10.1056/NEJMoa0805715
46. Horwood P, Bande G, Dagina R, Guillaumot L, Aaskov J, Pavlin B. The threat of chikungunya in Oceania. *Western Pac Surveill Response J*. 2013;4: 8–10. doi:10.5365/WPSAR.2013.4.2.003
47. Musso D, Nhan T, Robin E, Roche C, Bierlaire D, Zisou K, et al. Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. *Euro Surveill*. 2014;19. doi:10.2807/1560-7917.es2014.19.14.20761
48. WHO WHO. Zika Virus [Internet]. 20 Jul 2018 [cited 16 Dec 2020]. Available: <https://www.who.int/en/news-room/fact-sheets/detail/zika-virus>
49. Lo Presti A, Ciccozzi M, Cella E, Lai A, Simonetti FR, Galli M, et al. Origin, evolution, and phylogeography of recent epidemic CHIKV strains. *Infect Genet Evol*. 2012;12: 392–398. doi:10.1016/j.meegid.2011.12.015
50. Ross RW. The Newala epidemic. III. The virus: isolation, pathogenic properties and relationship to the epidemic. *J Hyg (Lond)*. 1956;54: 177–191. doi:10.1017/S0022172400044442
51. Weaver SC. Arrival of chikungunya virus in the new world: prospects for spread and impact on public health. *PLoS Negl Trop Dis*. 2014;8: e2921. doi:10.1371/journal.pntd.0002921

52. Robinson MC. An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953. *Trans R Soc Trop Med Hyg.* 1955;49: 28–32. doi:10.1016/0035-9203(55)90080-8
53. Powers AM, Logue CH. Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. *J Gen Virol.* 2007;88: 2363–2377. doi:10.1099/vir.0.82858-0
54. WHO WHO. Chikungunya. In: WHO | World Health Organization [Internet]. 15 Sep 2020 [cited 29 Dec 2020]. Available: <https://www.who.int/news-room/fact-sheets/detail/chikungunya>
55. Cassadou S, Boucau S, Petit-Sinturel M, Huc P, Leparc-Goffart I, Ledrans M. Emergence of chikungunya fever on the French side of Saint Martin island, October to December 2013. *Euro Surveill.* 2014;19. doi:10.2807/1560-7917.es2014.19.13.20752
56. Ramon-Pardo P, Cibrelus L, Yactayob S. Chikungunya: case definitions for acute, atypical and chronic cases [Internet]. 14 Aug 2015 [cited 5 Feb 2021]. Available: https://apps.who.int/iris/bitstream/handle/10665/242406/WER9033_410-414.PDF?sequence=1
57. Volk SM, Chen R, Tsetsarkin KA, Adams AP, Garcia TI, Sall AA, et al. Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. *J Virol.* 2010;84: 6497–6504. doi:10.1128/JVI.01603-09
58. Alanagreh L, Alzoughool F, Atoum M. The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms. *Pathogens.* 2020;9. doi:10.3390/pathogens9050331
59. Malik YA. Properties of Coronavirus and SARS-CoV-2. *Malays J Pathol.* 2020;42: 3–11.
60. Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect.* 2020;9: 221–236. doi:10.1080/22221751.2020.1719902
61. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene Reports.* 2020;19: 100682. doi:10.1016/j.genrep.2020.100682
62. Hamre D, Procknow JJ. A new virus isolated from the human respiratory tract. *Proc Soc Exp Biol Med.* 1966;121: 190–193. doi:10.3181/00379727-121-30734
63. McIntosh K, Dees JH, Becker WB, Kapikian AZ, Chanock RM. Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. *Proc Natl Acad Sci USA.* 1967;57: 933–940. doi:10.1073/pnas.57.4.933
64. van der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJM, Wolthers KC, et al. Identification of a new human coronavirus. *Nat Med.* 2004;10: 368–373. doi:10.1038/nm1024
65. Woo PCY, Lau SKP, Chu C, Chan K, Tsoi H, Huang Y, et al. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol.* 2005;79: 884–895. doi:10.1128/JVI.79.2.884-895.2005
66. Chafekar A, Fielding BC. MERS-CoV: Understanding the Latest Human Coronavirus Threat. *Viruses.* 2018;10. doi:10.3390/v10020093
67. de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, et al. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *J Virol.* 2013;87: 7790–7792. doi:10.1128/JVI.01244-13
68. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol.* 2009;7: 439–450. doi:10.1038/nrmicro2147
69. Chen L, Zhong L. Genomics functional analysis and drug screening of SARS-CoV-2. *Genes Dis.* 2020;7: 542–550. doi:10.1016/j.gendis.2020.04.002

70. Johnson HC, Gossner CM, Colzani E, Kinsman J, Alexakis L, Beauté J, et al. Potential scenarios for the progression of a COVID-19 epidemic in the European Union and the European Economic Area, March 2020. *Euro Surveill.* 2020;25. doi:10.2807/1560-7917.ES.2020.25.9.2000202
71. Chen J. Pathogenicity and transmissibility of 2019-nCoV-A quick overview and comparison with other emerging viruses. *Microbes Infect.* 2020;22: 69–71. doi:10.1016/j.micinf.2020.01.004
72. Bedford J, Enria D, Giesecke J, Heymann DL, Ihekweazu C, Kobinger G, et al. COVID-19: towards controlling of a pandemic. *Lancet.* 2020;395: 1015–1018. doi:10.1016/S0140-6736(20)30673-5
73. WHO WHO. Global Situation of COVID-19 [Internet]. [cited 16 Dec 2021]. Available: <https://covid19.who.int>
74. WHO WHO. Nota Técnica: Caracterização genômica de SARS-CoV-2 e variantes circulantes na região das Américas [Internet]. 8 Oct 2020 [cited 26 Feb 2021]. Available: <https://www.paho.org/pt/documentos/nota-tecnica-caracterizacao-genomica-sars-cov-2-e-variantes-circulantes-na-regiao-das>
75. Costa GS, Cota W, Ferreira SC. Outbreak diversity in epidemic waves propagating through distinct geographical scales. *Phys Rev Research.* 2020;2: 043306. doi:10.1103/PhysRevResearch.2.043306
76. Silva SJR da, Pena L. Collapse of the public health system and the emergence of new variants during the second wave of the COVID-19 pandemic in Brazil. *One Health.* 2021;13: 100287. doi:10.1016/j.onehlt.2021.100287
77. Walker PGT, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science.* 2020;369: 413–422. doi:10.1126/science.abc0035
78. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324
79. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics.* 2019;35: 871–873. doi:10.1093/bioinformatics/bty695
80. de Oliveira EC, Fonseca V, Xavier J, Adelino T, Morales Claro I, Fabri A, et al. Short report: Introduction of chikungunya virus ECSA genotype into the Brazilian Midwest and its dispersion through the Americas. *PLoS Negl Trop Dis.* 2021;15: e0009290. doi:10.1371/journal.pntd.0009290
81. Iani FCM, Giovanetti M, Fonseca V, Souza WM, Adelino TER, Xavier J, et al. Epidemiology and evolution of Zika virus in Minas Gerais, Southeast Brazil. *Infect Genet Evol.* 2021;91: 104785. doi:10.1016/j.meegid.2021.104785
82. Adelino TÉR, Giovanetti M, Fonseca V, Xavier J, de Abreu ÁS, do Nascimento VA, et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. *Nat Commun.* 2021;12: 2296. doi:10.1038/s41467-021-22607-0
83. Xavier J, Giovanetti M, Adelino T, Fonseca V, Barbosa da Costa AV, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg Microbes Infect.* 2020;9: 1824–1834. doi:10.1080/22221751.2020.1803146
84. Giandhari J, Pillay S, Wilkinson E, Tegally H, Sinayskiy I, Schuld M, et al. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int J Infect Dis.* 2021;103: 234–241. doi:10.1016/j.ijid.2020.11.128
85. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen

- novel lineages of SARS-CoV-2 in South Africa. *Nat Med*. 2021;27: 440–446. doi:10.1038/s41591-021-01255-3
86. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. 2021;592: 438–443. doi:10.1038/s41586-021-03402-9
 87. Xavier J, Giovanetti M, Fonseca V, Thézé J, Gräf T, Fabri A, et al. Circulation of chikungunya virus East/Central/South African lineage in Rio de Janeiro, Brazil. *PLoS One*. 2019;14: e0217871. doi:10.1371/journal.pone.0217871
 88. Gräf T, Vazquez C, Giovanetti M, de Bruycker-Nogueira F, Fonseca V, Claro IM, et al. Epidemiologic history and genetic diversity origins of chikungunya and dengue viruses, Paraguay. *Emerging Infect Dis*. 2021;27: 1393–1404. doi:10.3201/eid2705.204244
 89. Josseran L, Paquet C, Zehgnoun A, Caillere N, Le Tertre A, Solet J-L, et al. Chikungunya disease outbreak, Reunion Island. *Emerging Infect Dis*. 2006;12: 1994–1995. doi:10.3201/eid1212.060710
 90. Tabachnick WJ. Nature, nurture and evolution of intra-species variation in mosquito arbovirus transmission competence. *Int J Environ Res Public Health*. 2013;10: 249–277. doi:10.3390/ijerph10010249
 91. Ch O, Rosa APD, Tang AT, Amaral R, Afonso Dinis Costa Passos, Tauil PL. Surto de dengue em Boa Vista, Roraima Nota previa. *Revista Do Instituto De Medicina Tropical De Sao Paulo*. 1983;25: 53–54.
 92. OPAS / OMS. Reported cases of dengue fever in the Americas. [Internet]. [cited 1 Jun 2021]. Available: <https://www.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html>
 93. Murhekar MV, Manickam P, Kumar RM, Ganesakumar SRB, Ramachandran V, Ramakrishnan R, et al. Treatment practices & laboratory investigations during chikungunya outbreaks in South India. *Indian J Med Res*. 2011;133: 546–547.
 94. Van Bortel W, Dorleans F, Rosine J, Blateau A, Rousset D, Matheus S, et al. Chikungunya outbreak in the Caribbean region, December 2013 to March 2014, and the significance for Europe. *Euro Surveill*. 2014;19. doi:10.2807/1560-7917.es2014.19.13.20759
 95. Vega-Rúa A, Lourenço-de-Oliveira R, Mousson L, Vazeille M, Fuchs S, Yébakima A, et al. Chikungunya virus transmission potential by local Aedes mosquitoes in the Americas and Europe. *PLoS Negl Trop Dis*. 2015;9: e0003780. doi:10.1371/journal.pntd.0003780
 96. Akhrymuk I, Kulemzin SV, Frolova EI. Evasion of the innate immune response: the Old World alphavirus nsP2 protein induces rapid degradation of Rpb1, a catalytic subunit of RNA polymerase II. *J Virol*. 2012;86: 7180–7191. doi:10.1128/JVI.00541-12
 97. Nunes MRT, Faria NR, de Vasconcelos JM, Golding N, Kraemer MUG, de Oliveira LF, et al. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med*. 2015;13: 102. doi:10.1186/s12916-015-0348-x
 98. Rodrigues Faria N, Lourenço J, Marques de Cerqueira E, Maia de Lima M, Pybus O, Carlos Junior Alcantara L. Epidemiology of Chikungunya Virus in Bahia, Brazil, 2014–2015. *PLoS Curr Influenza*. 2016;8. doi:10.1371/currents.outbreaks.c97507e3e48efb946401755d468c28b2
 99. de Jesus JG, Dutra KR, Sales FC da S, Claro IM, Terzian AC, Candido D da S, et al. Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. *Mem Inst Oswaldo Cruz*. 2020;115: e190423. doi:10.1590/0074-02760190423

100. Drumond BP, Mondini A, Schmidt DJ, Bosch I, Nogueira ML. Population dynamics of DENV-1 genotype V in Brazil is characterized by co-circulation and strain/lineage replacement. *Arch Virol.* 2012;157: 2061–2073. doi:10.1007/s00705-012-1393-9
101. Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. A single mutation in chikungunya virus affects vector specificity and epidemic potential. *PLoS Pathog.* 2007;3: e201. doi:10.1371/journal.ppat.0030201

APPENDICES

Appendix 1 Supplementary material to the manuscript entitled “Genome Detective: an automated system for virus identification from high-throughput sequencing data”

Supplementary Information:

Genome Detective: An Accurate, Fast and Automated System for Virus Identification from High-throughput next generation sequencing (NGS) data

1) Genome Detective NGS Pipeline Figure

2) Genome Detective NGS Reports

2A) Quality control (QC) and Filtering

2B) Summary table

2C) Detailed table

2D) Detailed Alignment Statistics and Mutations

2E) Mapping short reads to de novo assembled consensus or reference sequence

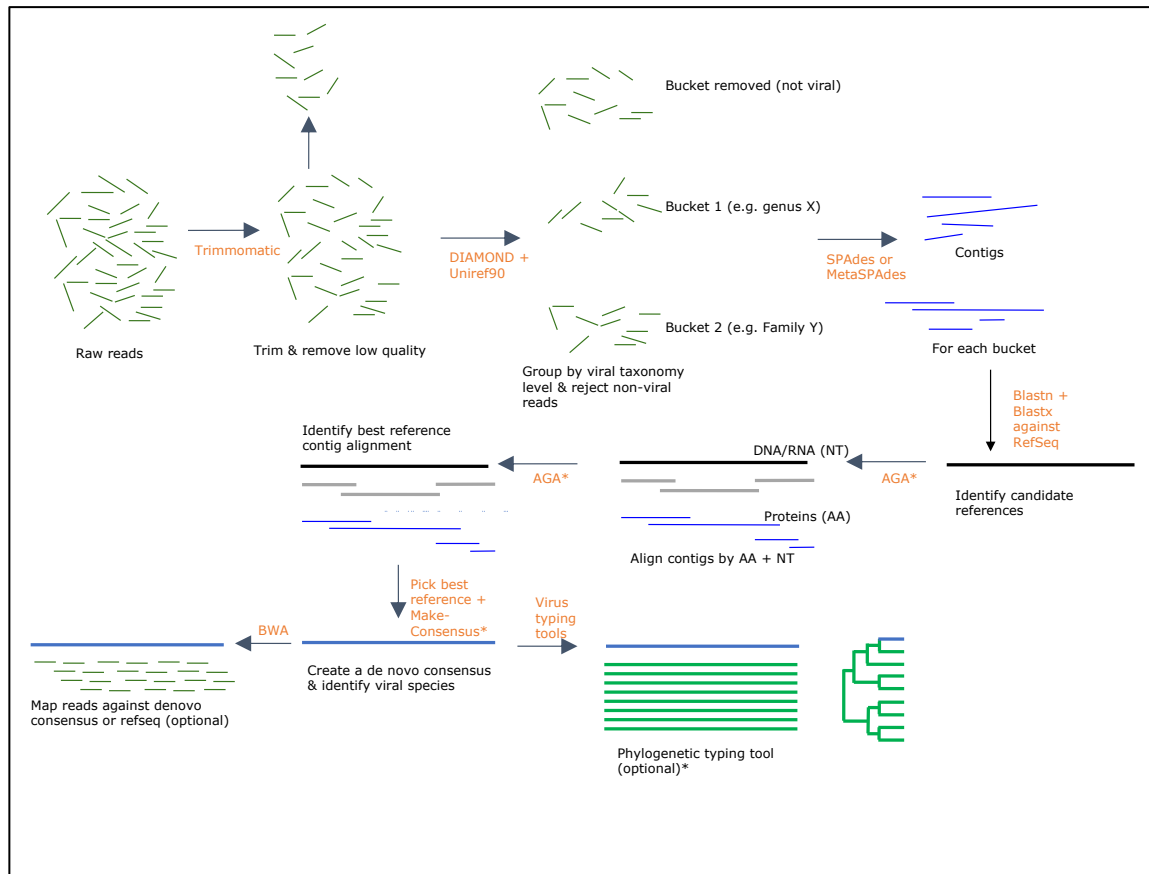
3) Accuracy: validation against published results

4) Performance: comparison competitor pipelines

5) References

1) Genome Detective NGS Pipeline

The diagram below shows the process of assembling next generation sequencing (NGS) short reads into *de novo* consensus sequences (Supplementary figure 1). Black text describes the process; orange text mentions the software applications used. SPAdes is used for single-ended reads and MetaSPAdes for paired-end reads. Green lines represent short reads and blue lines the contigs. Bold black lines represent virus reference sequences; grey lines the genomic proteins and green lines represent reference datasets used for phylogenetic typing.



Supplementary Figure 1: Schematic illustration of Genome Detective NGS Pipeline

2 - Genome Detective NGS Reports

2A) Quality control (QC) and Filtering:

The first part of the online report includes quality control (QC), filtering of viral reads and assembly and identification of viral taxonomy units (supplementary figure 1). The size of the input files, the original read length and the trimmed read length are summarized at the beginning of the report. This is followed by a report on the pre-processing and quality control steps, which filter low quality reads and reads that seem to be non-viral. The QC reports of the original submitted reads before and after pre-processing can be extracted from the report. The report also presents the total number of reads assembled and the computational time. A picture of the taxonomy units identified is created. The order of the taxonomy units is based on the number of the reads assembled in the table and in the taxonomy chart, with the largest number reported first.

You may bookmark this page to revisit results of this job (1242364539) later.

NGS ANALYSIS OF DRR049387

Size of input file(s)	82.39 MB	85.93 MB
Original read length	120	
Trimmed read length	50 - 105	
Submitted on	20-03-2018	

PREPROCESSING (0h 00m 58s)

Started with 2217376 reads, 88706 reads (4%) that did not pass qc, were removed.

Quality control (QC) reports

The preprocessing step will filter low quality reads and remove potential adapters. Below are QC reports of the original submitted reads and the reads after preprocessing.

Before preprocessing	QC report of reads 1	QC report of reads 2
After preprocessing	QC report of reads 1	QC report of reads 2

FILTERING (0h 02m 24s)

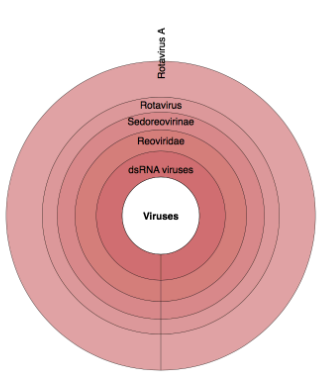
Started with 2128670 reads, 1872482 reads (87%) that did not appear to be viral, were removed ([Download](#)).

ASSEMBLY AND IDENTIFICATION (0h 16m 50s)

Started with 256188 reads.
Assembled: ~130000 reads.

Total computation time: 0h 20m 12s.






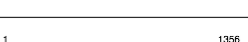

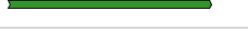



Host distribution Taxonomy chart Taxonomy tree
 Include discovery
 Scaling



Supplementary Figure 2: Genome Detective report provides information on the input file, quality control (QC) and filtering of viral reads and de novo assemblage.

2B) Summary table:

The summary table displays information on the assignment, number of *de novo* contigs, estimated number of reads, percentage (%) coverage of the genome, estimated mean depth coverage and nucleotide and amino acid identity (Supplementary table 3). The table also provides a link to download the *de novo* contigs and a link to download the detailed report on each assignment.


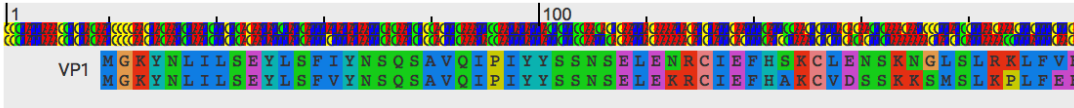
Assignment	# Contigs	Est. # Reads	Coverage (%)	Est. Depth of Coverage	Identity (%)		Report	Genome Coverage
					NT	AA		
Rotavirus A (segment 1)	1	~20000	100	~700	79	89.9	Report	
Rotavirus A (segment 2)	1	~18000	101.3	~600	79.5	91.4	Report	
Rotavirus A (segment 3)	1	~17000	100	~600	78.2	83.3	Report	
Rotavirus A (segment 4)	1	~16000	99.7	~700	70.9	71.8	Report	
Rotavirus A (segment 5)	1	~10000	97	~600	55	37.4	Report	
Rotavirus A (segment 6)	1	~9000	100	~700	79.1	92.7	Report	
Rotavirus A (segment 7)	1	~7000	95.9	~600	76.4	76.8	Report	
Rotavirus A (segment 9)	1	~6000	100	~600	76.3	81	Report	
Rotavirus A (segment 8)	1	~6000	98	~600	81.6	87.7	Report	
Rotavirus A (segment 11)	1	~4000	94.3	~700	88.7	91.4	Report	
Rotavirus A (segment 10)	1	~3000	99.9	~400	82.4	84.1	Report	

Download results: [XML File Table \(Excel format\)](#) [Table \(CSV format\)](#) [Contigs \(Fasta format\)](#) [BAM files](#)

Supplementary Figure 3: Genome Detective summary table report provides information on the assignment, number of *de novo* contigs, estimated number of reads, percentage (%) coverage of the genome, estimated mean depth coverage and nucleotide and amino acid identity.

2C) Detailed table:

A detailed table of the assembly is provided for each assignment. This table presents details on the assembly, assignment, alignment, genome region and codon alignment (Supplementary Figure 4).

NGS DETAILS	
ASSEMBLY	
Coverage length	3302 (1 contig(s))
Est. depth of coverage	~700
Est. number of reads	~20000
Ambiguities	0
COVERAGE DETAILS..	
ASSIGNMENT	
Type	Rotavirus A (Taxonomy ID: 28875)
Reference Genome	NC_011507.2 (Length: 3302bp)
Host(s)	Homo sapiens / Chlorocebus pygerythrus (host info)
NT Identity (%)	79.0127
AA Identity (%)	89.899
Number of stop codons	1
Number of CDS	1
ALIGNMENT	
Segment	1
Alignment score	3832 (NT) + 6517 (AA) = 10349
Concordance (%)	75.3916
Alignment method	Global, optimal, nucleotide + amino acids (AGA)
NT Alignment	Download alignment (FASTA)
CDS Alignments	Download CDS alignments (FASTA)
Contigs Alignments	Download contigs alignment (FASTA)
GENOME REGION	
Sequence starts at position 1 and ends at position 3302 relative to NC_011507.2 reference sequence.	
	
	
ALIGNMENT DETAILED STATISTICS	

Supplementary Figure 4: Genome Detective detailed table report for Rotavirus A de novo whole genome. This report also contains a diagram of the whole genome and the alignment.

2D) Detailed Alignment Statistics and Mutations:

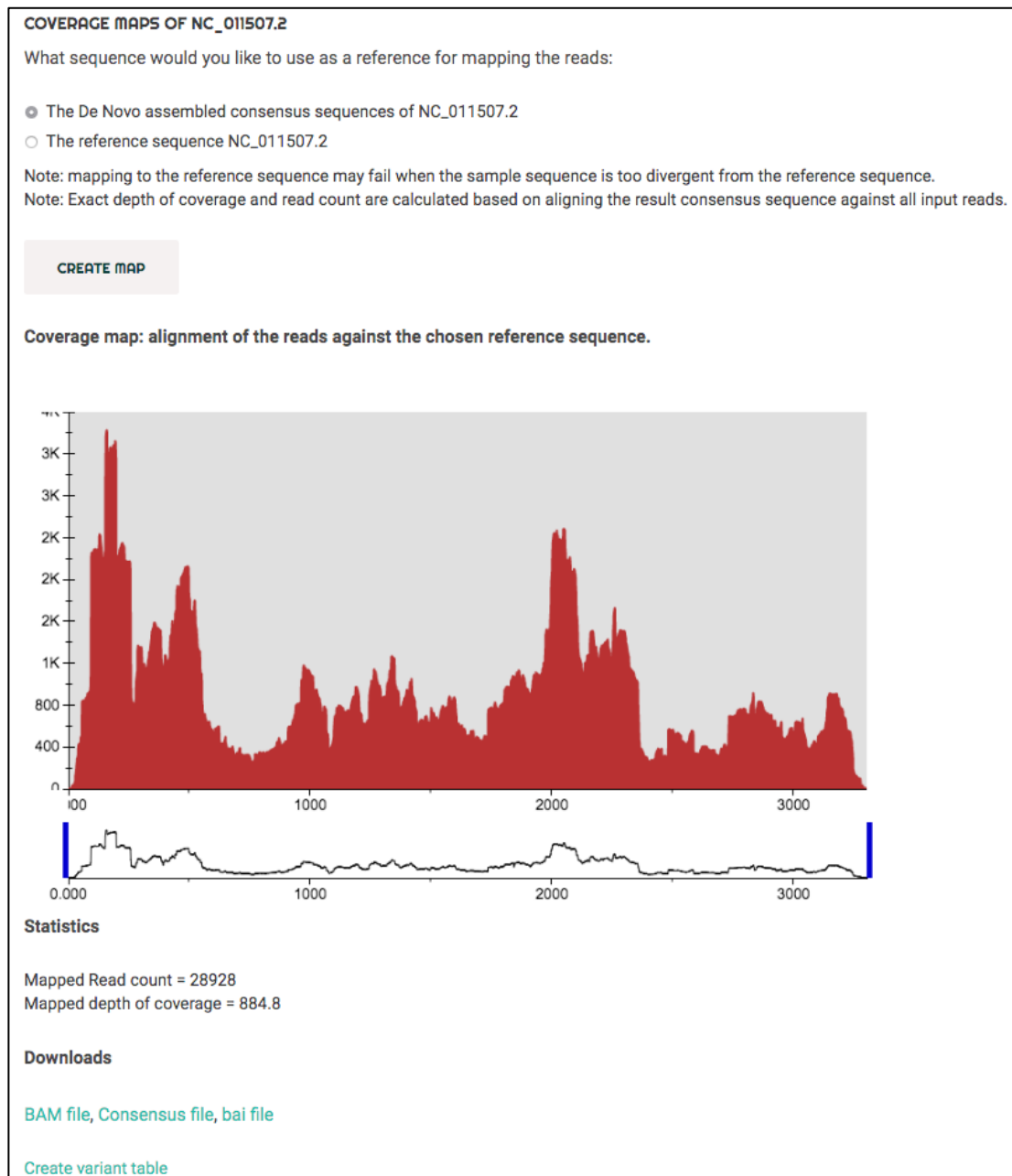
The detailed alignment statistics table below shows the nucleotide positions, coding regions and proteins (Supplementary Figure 5). All mutations in the protein are also presented with the following nomenclature (example: I15V, “I” is the wild type amino acid, “15” is the codon position and “V” is the mutant amino acid).

ALIGNMENT DETAILED STATISTICS												
	begin	end	coverage	score	concordance	matches	identities	inserts	deletes	misalignments	frameshifts	stop codons
NT	1	3302	100%	3832	58%	3302 (100%)	2609 (79%)	0	0			
CDS												
1_VP1	1	1089	100%	6517	91%	1089 (100%)	979 (90%)	0	0	0	0	1
Proteins	HIDE MUTATIONS											
VP1 (YP_0023...	1	1089	100%	6517	91%	1089 (100%)	979 (90%)	0	0	0	0	1
I15V N36K S43A L46V E47D N48S N51K G52S L53M R56K K57P V60E N63K E67D K91N V104A P120Q T121S K151R E153S E158D K159A L160N V162I A187E D194E D221E V231I V255I N273D R277K T284I N289Q Q290E L292I R296K A298V K305R D309E R314C I316L P320T A323S L340I N354D D357G R361N V371I E377V M428I N430H E431G I437V K445R I482S N512S T514S I544M A546S A552P L555I S616A H619Y A622I E642D I648V E657H T658I A660S R721K M737I A804S S806Q T813N R814K E817D F821I N825V R829K L833V I845V S846Y T866S S870N K882R T886V V887T D889E H891N P893S Q897R I910V D924S A929S R932K H954R N956Q I963V I967V K969P I970V D973G K979R R987K K1016R V1041I Y1044H S1050A S1075A N1082S												
PHYLOGENETIC ANALYSIS												
Sub-typing tool for Rotavirus A is not available.												

Supplementary Figure 5: Genome Detective detailed alignment statistics and mutations.

2E) Mapping short reads to de novo assembled consensus or reference sequence:

Genome detective allows the short reads to be mapped to the de novo assembly or to the viral reference sequence from NCBI RefSeq (Supplementary figure 6). Genome detective allows the short read assembly to be downloaded as BAM files and a single nucleotide position (SNP) variant table to be created.



Supplementary Figure 6: Genome Detective coverage maps can be created with the de-novo assembled consensus or with the reference sequence for a given virus. In this example, Rotavirus segment 1 is fully sequenced and assembled.

3) Accuracy: validation against published results

3A) Detailed results are presented in Supplementary Table 1, which is an Excel table of the supplementary information

In order to validate the results of Genome Detective, we compared them with the published results from 208 datasets. Summary results are presented in Table 1 of the manuscript. Detailed results are presented in Supplementary Table 1, an Excel table which is part of the supplementary information. Supplementary Table 1 contains detailed information about the 208 datasets used in the validation of Genome Detective. Variables are explained below.

Dataset name	Run accession number (SRR/ERR) of the dataset
Reference genome	RefSeq accession number of the closest related virus as identified by Genome Detective
Assigned species	Virus species identified by Genome Detective
Assigned genus	Virus genus identified by Genome Detective
Assignment agreement	Agreement with published result ('TRUE or FALSE')
Contigs count	The number of contigs generated
Coverage %	The percentage of the reference genome covered by the contigs
Reads count	The total number of reads used in the generation of the contigs
Deep coverage	The average deep coverage of the contigs, as determined by the function: $read\ count * read\ length / contig\ length$, as estimated by SPAdes
N50 (Assembly quality)	A statistical measure of the average length of the contigs. It is widely used to judge assembly quality based on contig lengths
Assignment quality	
The next 5 parameters compare detected virus sequence to the reference genome	
AA identity	The percentage amino acid identity in all coding regions of the detected virus as determined by the function: $Total\ number\ of\ matching\ amino\ acids\ in\ coding\ regions\ alignment / coding\ regions\ alignment\ length * 100$
AA quality	$Amino\ acid\ alignment\ score\ (see\ AGA\ for\ detailed\ scoring) / amino\ acid\ sequence\ length$
NT identity	The percentage nucleotide identity of the detected virus, as determined by the function: $Number\ of\ matching\ nucleotides\ in\ alignment / alignment\ length * 100$
NT quality	$Nucleotide\ alignment\ score\ (see\ AGA\ for\ detailed\ scoring) / nucleotide\ sequence\ length$
Frame shifts	The number of frame shifts detected in all coding regions of the detected virus
Stop codon (in CDs)	The total number of stop codons within coding regions of the detected virus sequence
Ambiguities (NT)	The number of nucleotide ambiguities in the detected virus sequence
Ambiguities (in CDs)	The number of amino acid ambiguities in the detected virus coding regions
Assignment quality	“GOOD” indicates that the assignment is considered reliable based on the heuristic of having sufficient nucleotide identity for a sufficient part of the genome, as determined by the function: $“NT\ identity” / 100 * “Coverage” + (“Reference\ length” - “Coverage”) * 0.45 > 0.5 * “Reference\ length”$
Sample	
General information about the sample	
Time	The total analysis time
# Reads before QC	The read count in the original dataset
# Reads after QC	The number of reads that remain after preprocessing (trimming and removing low quality reads)
# Reads after filtering	The number of reads that remain after filtering (alignment against UniRef90 protein database)
Read length before QC	The average read length in the original dataset
Read length after QC	The average read length after preprocessing

Input file (zipped) size	The dataset size
--------------------------	------------------

Supplementary Table 1 legend for Excel table showing the results of the evaluation of the 208 datasets.

3B) Detailed information on the seven datasets presented in Table 1 of the manuscript.

As previously mentioned in the manuscript and Table 1, 208 datasets from eight studies were used in the validation of Genome Detective. Below, we provide information on the datasets used in the validation process. In addition, we mention the tables and figures from the original publications that were used to validate our results.

Publication 1 - Virome synthetic datasets.

We first validated Genome Detective by using synthetic virus datasets originally prepared to optimise laboratory-based virus extraction procedures [Conceição-Neto et al. 2015]. Viruses were carefully selected to cover the range of naturally occurring diversity. This published dataset also included carefully validated quantitative results, confirmed with quantitative PCR. The Supplementary Table S4 in the article described the exact quantity of each virus in every sample. All samples contained Circovirus, Parvovirus, Polyomavirus, Pepino mosaic virus, Rotavirus, Coronavirus, Herpesvirus and Mimivirus in different quantities. Genome Detective identified all of the 64 viruses in the synthetic dataset was reconstructed and for seven the partial genome.

Sample	Reconstructed	Also detected
SRR3458562	Identified 8/8	Koala retrovirus
SRR3458563	Identified 7/8 (Herpesvirus detected with low genome coverage)	Koala retrovirus
SRR3458564	Identified 7/8 (Mimivirus detected with low genome coverage)	Koala retrovirus and Baboon endogenous virus strain M7
SRR3458565	Identified 7/8 (Mimivirus detected with low genome coverage)	Koala retrovirus
SRR3458566	Identified 7/8 (Mimivirus detected with low genome coverage)	Koala retrovirus
SRR3458567	Identified 7/8 (Mimivirus detected with low genome coverage)	Koala retrovirus
SRR3458568	Identified 7/8 (Mimivirus detected with low genome coverage)	Gibbon ape leukemia virus
SRR3458569	Identified 7/8 (Mimivirus detected with low genome coverage)	Koala retrovirus

Notes:
 In all cases, all Rotavirus segments were identified
 In several samples, Herpesvirus or Mimivirus were detected with very low genome coverage (< 0.05 %) due to low concentration and large genome size
 BLAST search revealed that the additional (false positive) viruses in the “Also detected” column were most likely endogenous viruses that were not in RefSeq, and Genome Detective assigned them to the closest available reference genome

Supplementary Table 2: Detailed results from synthetic (metagenomic) datasets

Publications 2-6 - Single virus amplicon-based datasets – HIV, RSV, Rotavirus, Norovirus, Influenza A and MERs.

Single virus amplicon-based datasets were used to validate the pipeline with specific viruses [Agoti et al. 2015, Cotton et al. 2014, Rutvisuttinunt et al. 2015, Cotton et al. 2013, de Oliveira et al. 2018]

These included single viruses with segmented genomes (RSV, Rotavirus, Norovirus, Influenza A, MERS) and unsegmented genomes (HIV) from amplicon-based NGS. Because the study design included amplification of the viruses of interest, the reads were expected to cover the amplified fragments. In the single virus datasets, we reconstructed the genomes of all of the amplified virus fragments with high accuracy. Furthermore, Genome Detective produced longer contigs than most of the competitor pipelines.

Publication 7 – Human and pig samples for Rotavirus.

This study analyzed human and pig samples for Rotavirus (RV). They used a primer-independent, agnostic, deep sequencing approach [Phan et al. 2016].

Publication 8 – Metagenomic dataset

The study [Cotten et al. 2014] analyzed 20 metagenomic datasets and they provided a heat map (Fig 6 of the article). Genome Detective identified the 66 virus species with more than 10 reads as the original study did.

4) Performance: comparison with competitor pipelines

In order to compare Genome Detective's performance with other pipelines, we used four published datasets: SRR1170797, SRR1106548, DRR049387 and ERR690519. These are four of the five datasets that the authors of drVM [Lin & Liao et al. 2017] used to compare with three other tools, SURPI [Naccache et al. 2014], VIP [Li et al. 2016] and VirusTap [Yamashita et al. 2016] (Supplementary Table 3).

We found that, in general, Genome Detective created longer, more accurate contigs than drVM, SURPI, VIP and VirusTap. In addition, Genome Detective was faster than the four other tools. For example, in the HIV-1 dataset (SRR1106548), Genome Detective assembled a near complete genome (8,334 of 9,181 bps) in 430 sec, whereas drVM identified a 3,055 bp contig in 608 sec. Both Genome Detective and drVM also identified Torque teno virus. For the Rotavirus dataset (DRR049387), Genome Detective identified all of the 11 segments of Rotavirus A (segment 1 to 11), each with one contig covering 97-100% of the segment, whereas drVM identified 13 contigs covering only seven segments. The time for this run in Genome Detective was 440 seconds whereas it took 464 seconds in drVM [Lin & Liao 2017]. For Influenza A virus (ERR690519), we identified the same eight segments as drVM in less than half the time. Supplementary table 3 is adapted from the drVM paper, which compared performance between drVM, SURPI, VIP and VirusTAP. We were unable to locate the fifth dataset (SRR062073) in the public databases. (see <https://www.ebi.ac.uk/ena/data/view/SRR062073&display=html>).

Target virus (run accession)	Read bases (Mbp)		Genome Detective	drVM	SURPI (comprehensive)	VIP (sense)	VirusTAP
Bovine viral diarrhea virus (SRR1170797)	12.5	Run time	71 s	149s	52 365 s	1 683 s	71 s
		Result	11 906 bp	12 224 bp	262 bp	9078 bp	353 bp
Human immunodeficiency virus (SRR1106548)	600.9	Run time	430 s	598 s	32 604 s	25 049 s	1 388 s
		Result	8 334 bp	3 055 bp	799 bp	4632 bp	2 896 bp
Human rotavirus A (DRR049387)	266.1	Run time	440 s	464 s	59 510 s	6259 s	925 s
		Result	11 contigs (11 segments, 99% complete)	13 contigs (7 segments)	13 contigs	41 377 reads	11 contigs
Influenza A virus (ERR690519)	3300	Run time	1 423 s	12 697 s	16 997 s	86 001 s	4 504 s

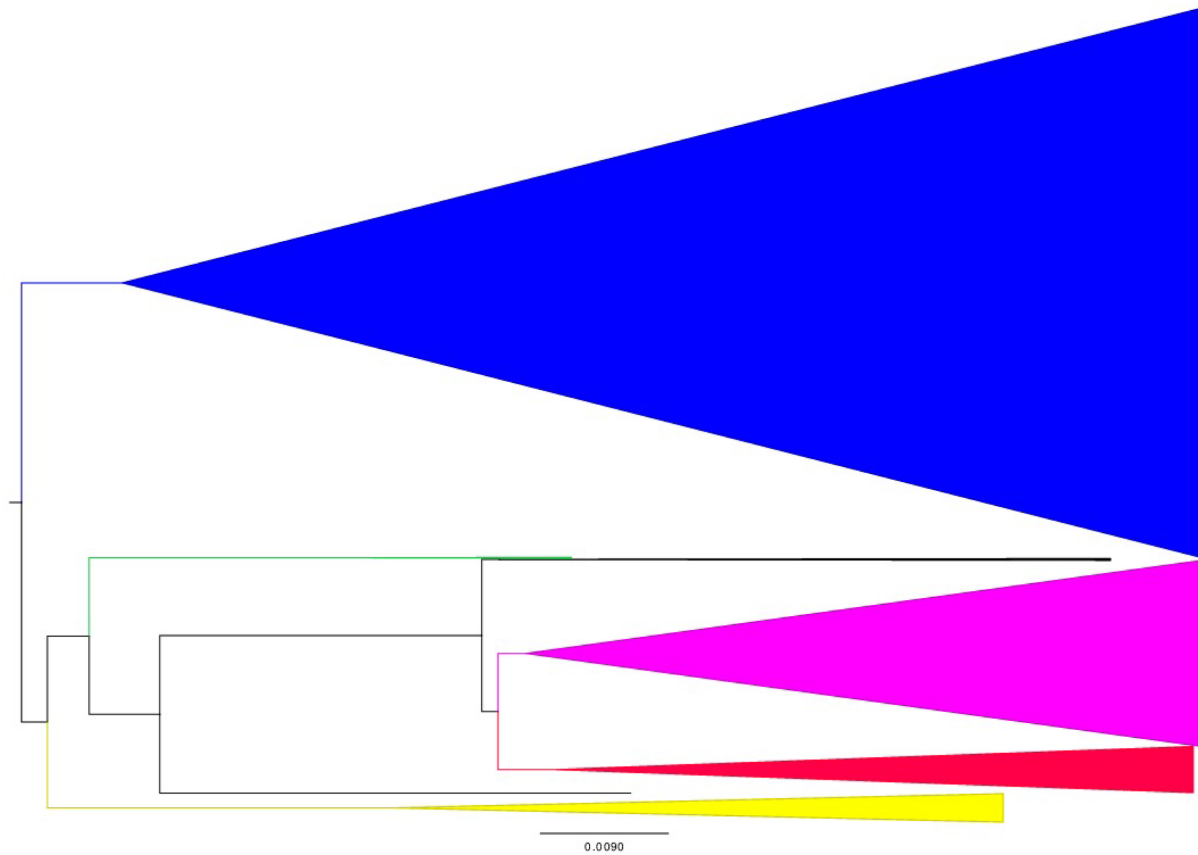
		Result	8 contigs (8 segments, 99% complete)	8 segments	11 contigs	2673 reads	34 contigs
--	--	--------	--------------------------------------	------------	------------	------------	------------

Supplementary Table 3: Performance Comparison with other pipelines. We analyzed the Genome Detective data for the same datasets analyzed by drVM. This table was adapted from Lin & Liao 2017 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466706/table/tbl3/>.) Genome Detective's analysis was executed on a quad-core CPU with 64 GB RAM. The drVM, SURPI and VIP analyses were executed on a quad-core CPU with 128 GB RAM, and the VirusTAP analyses executed on a 120-core CPU with 1 TB RAM.

5) References:

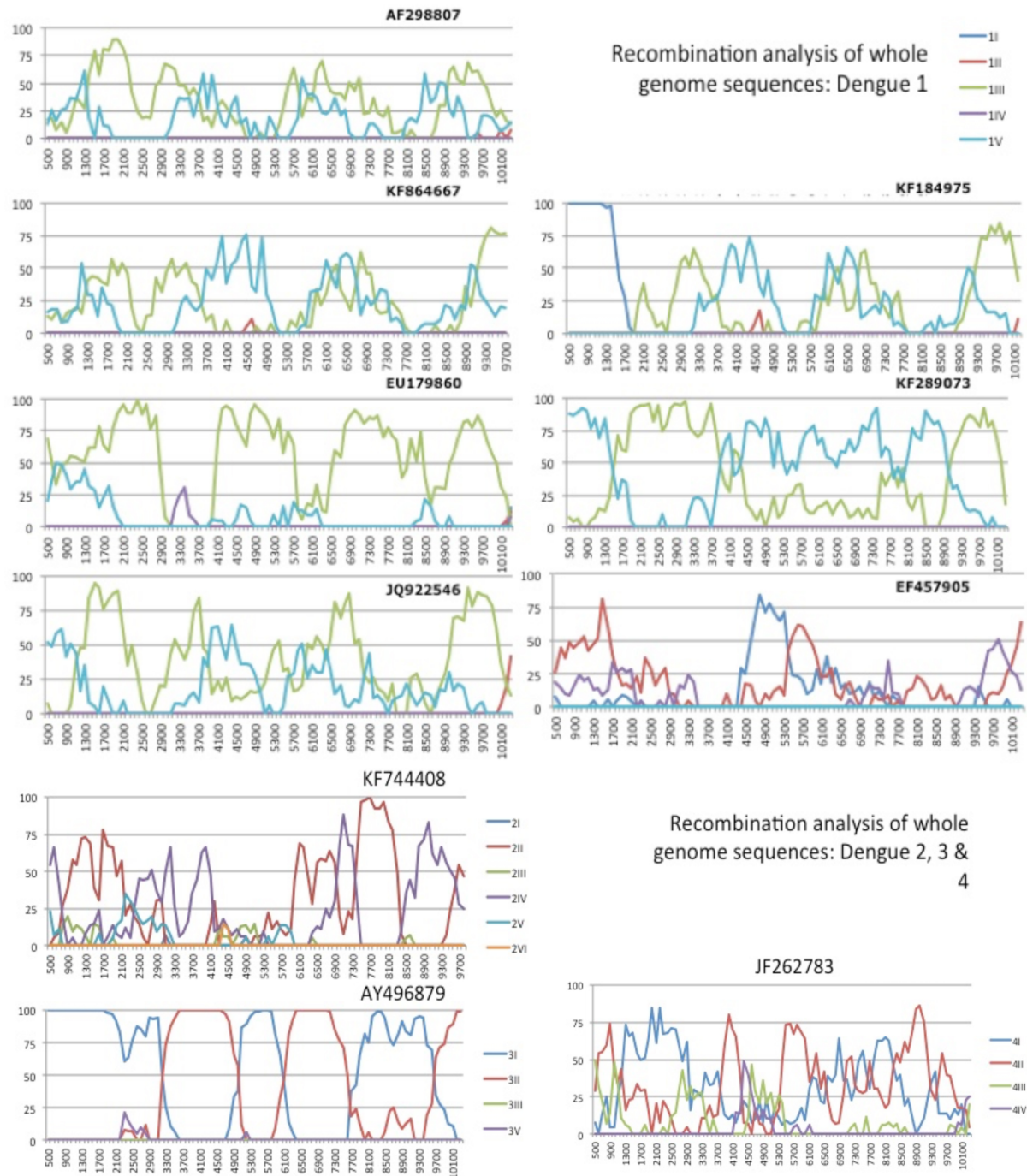
- Agoti, C. N., Otieno, J. R., Munywoki, P. K., Mwihuri, A. G., Cane, P. A., Nokes, D. J., ... Cotten, M. (2015). Local Evolutionary Patterns of Human Respiratory Syncytial Virus Derived from Whole-Genome Sequencing. *Journal of Virology*, 89(7), 3444–3454. <http://doi.org/10.1128/JVI.03391-14>
- Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., ... Matthijnsens, J. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific Reports*, 5, 16532. Retrieved from <http://dx.doi.org/10.1038/>
- Cotten, M., Petrova, V., Phan, M. V. T., Rabaa, M. A., Watson, S. J., Ong, S. H., ... Baker, S. (2014). Deep Sequencing of Norovirus Genomes Defines Evolutionary Patterns in an Urban Tropical Setting. *Journal of Virology*, 88(19), 11056–11069. <http://doi.org/10.1128/JVI.01333-14>
- [Cotten, M., Oude Munnink, B., Canuti, M., Deijs, M., Watson, S. J., Kellam, P., & van der Hoek, L. \(2014\). Full Genome Virus Detection in Fecal Samples Using Sensitive Nucleic Acid Preparation, Deep Sequencing, and a Novel Iterative Sequence Classification Algorithm. PLoS ONE, 9\(4\), e93269. http://doi.org/10.1371/journal.pone.0093269](http://doi.org/10.1371/journal.pone.0093269)
- Cotten M, Watson SJ, Kellam P, Al-Rabeeh AA, Makhdoom HQ, Assiri A, ... Memish ZA. (2013). Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet* 382(9909):1993-2002. doi: 10.1016/S0140-6736(13)61887-5.
- de Oliveira T, Giandhari J. HIV-1 – unpublished, but data deposited at SRA - PRJNA434385.
- Lin, H.-H., & Liao, Y.-C. (2017). drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. *GigaScience*, 6(2), 1–10. Retrieved from <http://dx.doi.org/10.1093/>
- Phan, M. V. T., Anh, P. H., Cuong, N. Van, Munnink, B. B. O., van der Hoek, L., ... My, P. T. (2016). Unbiased whole-genome deep sequencing of human and porcine stool samples reveals circulation of multiple groups of rotaviruses and a putative zoonotic infection. *Virus Evolution*, 2(2), vew027-vew027. Retrieved from <http://dx.doi.org/10.1093/ve/>
- Rutvisuttinunt, W., Chinnawirotpisan, P., Thaisomboonsuk, B., Rodpradit, P., Ajariyakhajorn, C., Manasatienkij, W., ... Fernandez, S. (2017). Viral subpopulation diversity in influenza virus isolates compared to clinical specimens. *Journal of Clinical Virology*, 68, 16–23. <http://doi.org/10.1016/j.jcv.>
- Wymant, C., Blanquart, F., Gall, A., Bakker, M., Bezemer, D., Croucher, N. J., ... Fraser, C. (2016). Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data. *bioRxiv*. Retrieved from <http://biorxiv.org/content/>

Appendix 2 Supplementary material to the manuscript entitled “A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes”



S1 Fig. Maximum likelihood phylogenetic tree of the DENV-sero1 outliers.

All full genome DENV-sero1 sequences were assigned to genotype-level using manual phylogenetic analysis and classification by the automated typing tool. In total, seven full genomes of DENV-sero1 could not be classified at genotype level by either classification method. These seven sequences are visualized in a phylogenetic tree of the WGS datasets, colored according to genotype. (I in blue, II in green, III in red, IV in yellow, V in pink) It can be seen that a divergent cluster of six genomes (AF298807, KF864667, KF184975, EU179860, KF289073 and JQ922546 in black) form an outlier clade and one genome (EF457905 in black) can be considered an outlier. However, note that these seven genomes could be properly assigned to serotype 1.



S2 Fig. Recombination analysis for the DENV whole genome sequences.

The bootscan results for the ten whole genomes of DENV that could not be classified at genotype level are shown. Boot-scanning analysis was performed using a window length of 1500 base pairs and a step size of 100 base pairs. The different colours represent the genotypes for each serotype. The X-axis represents the nucleotide position in the genome and the Y-axis represents bootstrap results in percentages. In total, 7 DENV-sero1 sequences were analysed and 1 sequence for each of the other serotypes, i.e. DENV-sero2, DENV-sero3 and DENV-sero4. We only found sequence AY496879 to be a recombinant of DENV genotype 3I and 3II. The other sequences are outliers (i.e. JF262783, KF744408, EF457905) or clades of outliers (i.e.: AF298807, KF864667 and KF184975 form an outlier clade; EU179860, KF289073 and JQ922546 form an outlier clade).

S1 Table. Reference strains selected for the DENV, CHIKV, ZIKV genotypes.

These reference sequences were selected to be representative for the diversity within the different DENV, CHIKV and ZIKV genotypes that circulate within these virus species.

ZIKV Genotype	Accession Number	Country	Year
African	AY632535	Uganda	1947
African	KF383116	Senegal	1968
African	HQ234500	Nigeria	1968
African	KF383115	Cent Afr Rep	1968
African	HQ234501	Senegal	1984
African	KF383117	Senegal	1997
African	KF383118	Senegal	2001
African	KF383119	Senegal	2001
African	KF383121	Senegal	-N/A-
African	KF268950	Cent Afr Rep	-N/A-
African	KF268949	Cent Afr Rep	-N/A-
Asian	HQ234499	Malaysia	1966
Asian	EU545988	Micronesia	2007
Asian	JN860885	Cambodia	2010
Asian	KF993678	Canada	2013
Asian	KJ776791	FrenchPolynesia	2013

CHIKV Genotype	Accession Number	Country	Year
Asian	HM045813	India	1963
Asian	EF027140	India	1963
Asian	EF027141	India	1973
Asian	HM045790	Philippines	1985
Asian	FN295483	Malaysia	2006
Asian	FJ807897	Taiwan	2007
ESCA_IOC	HM045811	Tanzania	1953
ESCA_IOC	HM045821	Senegal	1963
ESCA_IOC	AM258993	Reunion	2005
ESCA_IOC	AM258991	Seyche	2005
ECSA_IN	AB455494	Japan	2006
ESCA_IOC	EF012359	Mauri	2006
ECSA_IN	EU244823	Italy	2007
ECSA_IN	FJ445426	SriLanka	2008
ECSA_IN	FN295485	Malaysia	2008
ECSA_IN	GU199352	China	2008
ECSA_IN	GU301781	Thailand	2009
ESCA_IOC	HM045784	Central	-N/A-
ESCA_IOC	HM045822	Central	-N/A-
ESCA_IOC	HM045792	South	-N/A-
WestAfr	HM045786	Nigeria	1964
WestAfr	HM045785	Senegal	1966
WestAfr	HM045815	Senegal	1979
WestAfr	HM045817	Senegal	2005
WestAfr	HM045818	Ivory	-N/A-
WestAfr	HM045820	Ivory	-N/A-

Dengue Virus 1 Genotype	Accession Number	Country	Year
II	AF350498	-N/A-	1980
II	AY732478	Thailand	1991
II	AY732480	Thailand	1994
II	GQ868637	Cambodia	2000
II	AY732482	Thailand	2001
II	FJ469907	Singapore	2003
II	GQ199835	VietNam	2005
II	FJ176779	China	2006
II	AB608786	Taiwan	2008
II	GU131895	Cambodia	2009
II	HQ891316	Sri Lanka	2009
II	AY726552	Myanmar	2012
II	AF298808	Djibouti	-N/A-
III	AF180817	-N/A-	-N/A-
III	AY713473	Myanmar	1971
III	AY722801	Myanmar	1976
III	AY722802	Myanmar	1996
III	AY722803	Myanmar	1998
IIV	EF032590	-N/A-	1995
IIV	AB189121	Indonesia	1998

IIV	DQ672564	USA	2001
IIV	DQ672561	USA	2001
IIV	FJ196842	China	2003
IIV	GQ868602	Philippines	2004
IIV	AB204803	Japan	2004
IIV	JN697056	Malaysia	2005
IIV	U88535	-N/A-	-N/A-
IIV	AB074761	-N/A-	-N/A-
IIV	AB189120	Indonesia	-N/A-
1V	FJ205875	USA	1995
1V	AF311956	-N/A-	1997
1V	EU482567	USA	1998
1V	AB519681	Brazil	2001
1V	DQ285559	Reunion	2004
1V	HQ166037	Mexico	2008
1V	KJ189351	Puerto Rico	2012
1V	AY277666	-N/A-	-N/A-
1V	AY206457	-N/A-	-N/A-
1V	AY277665	-N/A-	-N/A-

Dengue Virus 2 Genotype	Accession Number	Country	Year
2I (American)	EU056812	Puerto Rico	1977
2I (American)	EU056811	Peru	1995
2I (American)	AF100469	-N/A-	-N/A-
2II (Cosmopolitan)	EU081180	Singapore	2005
2II (Cosmopolitan)	EU081179	Singapore	2005
2II (Cosmopolitan)	EU081177	Singapore	2005
2II (Cosmopolitan)	EU179859	Brunei	2006
2II (Cosmopolitan)	KC762660	Indonesia	2007
2II (Cosmopolitan)	KC762669	Indonesia	2007
2II (Cosmopolitan)	KC762680	Indonesia	2010
2II (Cosmopolitan)	KM279597	Singapore	2012
2III (SE Asian-America)	EU482582	USA	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	Dominican Republic	2001
2III (SE Asian-America)	FJ898461	Belize	2002
2III (SE Asian-America)	EU687216	USA	2005
2III (SE Asian-America)	EU687217	USA	2005
2III (SE Asian-America)	GQ199868	Nicaragua	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	Philippines	1995
2IV (Asian II)	KF744407	Philippines	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	Thailand	1974
2V (Asian I)	DQ181805	Thailand	1979
2V (Asian I)	DQ181804	Thailand	1984
2V (Asian I)	DQ181802	Thailand	1988
2V (Asian I)	GQ868545	-N/A-	1996
2V (Asian I)	DQ181798	Thailand	1999
2V (Asian I)	DQ181797	Thailand	2001
2V (Asian I)	FM210211	Viet Nam	2003
2V (Asian I)	GU131896	Cambodia	2007
2VI (Sylvatic)	EF105387	Nigeria	1966
2VI (Sylvatic)	EF105379	Malaysia	1970
2VI (Sylvatic)	EF105382	Burkina Faso	1980
2VI (Sylvatic)	EF105389	Senegal	1999
2VI (Sylvatic)	FJ467493	Malaysia	2008

Dengue Virus 3 Genotype	Accession Number	Country	Year
3I	AY858039	Indonesia	1998
3I	KC762682	Indonesia	2007
3I	KC762681	Indonesia	2007
3I	KC762686	Indonesia	2007
3I	KC762684	Indonesia	2007
3I	KC762687	Indonesia	2008
3I	KC762689	Indonesia	2008
3I	KC762688	Indonesia	2008
3I	KC762690	Indonesia	2008

3I	KC762685	Indonesia	2008
3I	KC762691	Indonesia	2008
3I	AY858037	Indonesia	-N/A-
3II	AY876494	Thailand	1994
3II	AY923865	Thailand	1994
3II	AY766104	Singapore	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	AY099337	Martinique	1999
3III	FJ639747	Venezuela	2000
3III	FJ547071	USA	2000
3III	FJ898458	Peru	2002
3III	AY679147	Brazil	2002
3III	EU529702	USA	2003
3III	FJ898442	Mexico	2007
3III	GQ868578	Colombia	2007
3III	AY099336	Sri Lanka	-N/A-
3III	AY770511	India	-N/A-
3V	EF629370	Brazil	2002
3V	AF317645	China	-N/A-

Dengue Virus 4 Genotype	Accession Number	Country	Year
4I	GQ868594	Philippines	1956
4I	AY618991	Thailand	1977
4I	FJ196850	China	1990
4I	AY618992	Thailand	2001
4I	KF041260	Pakistan	2009
4I	JQ513345	Brazil	2011
4II	GU289913	Colombia	1982
4II	JF262782	Haiti	1994
4II	AY762085	-N/A-	1995
4II	JF262781	Venezuela	1995
4II	GQ252675	USA	1995
4II	FJ024476	Colombia	1997
4II	FJ882581	Venezuela	2007
4II	JN983813	Brazil	2010
4II	AF326573	-N/A-	-N/A-
4III	AY618988	Thailand	1997
4III	AY618989	Thailand	1997
4IV	JF262780	Malaysia	1973
4IV	EF457906	Malaysia	1975

S2 Table. Phylogenetic signal estimated by likelihood mapping for DENV (DENV-sero1 to DENV-sero4), CHIKV and ZIKV sub-genomic regions.

Phylogenetic signal was calculated separately per protein by the likelihood mapping method implemented in the software TreePuzzle. Likelihood mapping analysis computes the likelihood of the three possible trees that can be constructed from all possible inter-genotype quartets of taxa. The results for the resolved quartets and unresolved quartets are shown in the table, while the partially resolved quartets are not listed (can be obtained by 100%—(un)resolved quartets). Partially resolved quartets represent the quartets for which conflicting phylogenetic signal or potential recombination is present. Genomic regions for which the percentage of resolved quartets is higher than 90% are shaded in orange and are considered to be characterized by sufficient phylogenetic signal.

	DENV-1		DENV-2		DENV-3		DENV-4	
	Resolved	Unresolved	Resolved	Unresolved	Resolved	Unresolved	Resolved	Unresolved
C	54.5%	42,3%	68,3%	25,8%	86,1%	9,8%	78,3%	15,5%
E	93.2%	3,3%	95,3%	2,5%	97,2%	1,3%	92,9%	3,4%
M	75.3%	20,2%	87,6%	8,8%	88,2%	7,3%	83,4%	12,2%
NS1	92.2%	4,5%	89,4%	8,1%	96,7%	1,3%	95,4%	2,3%
NS2A	82.5%	15,5%	87,7%	9,4%	88,5%	8,3%	87,3%	8,5%
NS2B	79.0%	19,2%	83,0%	14,4%	89,2%	6,3%	82,5%	12,7%
NS3	94.6%	2,8%	94,5%	2,6%	97,2%	1,3%	95,0%	2,7%
NS4A	73.8%	19,9%	78,1%	19,8%	88,7%	8,4%	84,7%	9,9%
NS4B	85.2%	12,7%	90,2%	5,7%	92,2%	4,1%	83,5%	10,6%
NS5	94.7%	2,0%	94,0%	3,2%	98,1%	0,7%	97,0%	1,4%

	ZIKV	
	Resolved	Unresolved
C	78.2%	3.9%
E	94.3%	1.5%
M	81.6%	4.8%
NS1	91.9%	1.8%
NS2A	93.5%	2.3%
NS2B	85.4%	1.1%
NS3	94.0%	1.7%
NS4A	90.8%	3.7%
NS4B	93.2%	2.0%
NS5	96.3%	2.4%

	CHIKV	
	Resolved	Unresolved
NSP1	52.2%	47.7%
NSP2	52.1%	47.9%
NSP3	62.3%	19.6%
NSP4	52.7%	46.8%
CAP	56.9%	35.6%
E3	58.2%	26.9%
E2*	77.1%	13.1%
E1	53.1%	41%

S3 Table. Evaluation of the automated phylogenetic method to classify DENV, CHIKV and ZIKV whole-genome genomes.

The new classification method consists of 2 parts: determining the species (and for DENV also the serotype) using a BLAST procedure, followed by determining the genotype using an automated phylogenetic method. Our method was able to assign all sequences in the whole-genome validation dataset to the right species and DENV serotype. Therefore, in this table, we focus on the classification performance with respect to genotype assignment, based on the output of the BLAST step (i.e. a dataset of the proper species and serotype). The classification results were compared to manual phylogenetic analysis. Column names: TP = total positives, TN = total negatives, FP = false positive, FN = false negative, SENS = sensitivity, SPEC = specificity, ACC = accuracy.

Virus species	Known	TP	TN	FP	FN	SENS	SPEC	ACC
Dengue Virus Serotype 1	1688	1688	3496	0	0	100,0%	100,0%	100,0%
i	1151	1151	4033	0	0	100,0%	100,0%	100,0%
ii	28	28	5156	0	0	100,0%	100,0%	100,0%
iii	26	26	5158	0	0	100,0%	100,0%	100,0%
iv	63	63	5121	0	0	100,0%	100,0%	100,0%
lv	413	413	4771	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	7							
Dengue Virus Serotype 2	1317	1317	3867	0	0	100,0%	100,0%	100,0%
2i (American)	47	47	5137	0	0	100,0%	100,0%	100,0%
2ii (Cosmopolitan)	245	245	4939	0	0	100,0%	100,0%	100,0%
2iii (SE Asian-America)	649	649	4535	0	0	100,0%	100,0%	100,0%
2iv (Asian II)	41	41	5143	0	0	100,0%	100,0%	100,0%
2v (Asian I)	317	317	4867	0	0	100,0%	100,0%	100,0%

2vi (Sylvatic)	17	17	5167	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							
Dengue Virus Serotype 3	897	897	4287	0	0	100,0%	100,0%	100,0%
3i	68	68	5116	0	0	100,0%	100,0%	100,0%
3ii	190	190	4994	0	0	100,0%	100,0%	100,0%
3iii	620	620	4564	0	0	100,0%	100,0%	100,0%
3v	18	18	5166	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							
Dengue Virus Serotype 4	216	216	4968	0	0	100,0%	100,0%	100,0%
4i	23	23	5161	0	0	100,0%	100,0%	100,0%
4ii	187	187	4997	0	0	100,0%	100,0%	100,0%
4iii	2	2	5182	0	0	100,0%	100,0%	100,0%
4iv	3	3	5181	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							
Total	4118							
Chikungunya Virus								
East-Central-South-African	274	274	4910	0	0	100,0%	100,0%	100,0%
Asian and Caribbean	364	364	4820	0	0	100,0%	100,0%	100,0%
West African	15	15	5169	0	0	100,0%	100,0%	100,0%
Total	653							
Zika Virus								
African	26	26	5158	0	0	100,0%	100,0%	100,0%
Asian	387	387	4797	0	0	100,0%	100,0%	100,0%
Total	413							
Total sequences	5184							

S4 Table. Evaluation of the automated phylogenetic method to classify DENV, CHIKV and ZIKV envelope genomes.

The new classification method consists of 2 parts: determining the species (and for DENV also the serotype) using a BLAST procedure, followed by determining the genotype using an automated phylogenetic method. Our method was able to assign all sequences in the envelope validation dataset to the right species and DENV serotype. Therefore, in this table, we focus on the classification performance with respect to genotype assignment, based on the output of the BLAST step (i.e. a dataset of the proper species and serotype). The classification results were compared to manual phylogenetic analysis. Column names: TP = total positives, TN = total negatives, FP = false positive, FN = false negative, SENS = sensitivity, SPEC = specificity, ACC = accuracy.

Virus species	Known	TP	TN	FP	FN	SENS	SPEC	ACC
Dengue Virus ENV								
Dengue Virus Serotype 1	1688	1688	5374	0	0	100,0%	100,0%	100,0%
DQ859064	1151	1148	5911	0	3	99,7%	100,0%	100,0%
1ii	28	28	7034	0	0	100,0%	100,0%	100,0%
1iii	26	25	7034	2	1	96,2%	100,0%	100,0%
1iv	63	59	6998	1	4	93,7%	100,0%	99,9%
1v	413	413	6649	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	7							
Dengue Virus Serotype 2	1317	1317	5745	0	0	100,0%	100,0%	100,0%
2i (American)	47	47	7015	0	0	100,0%	100,0%	100,0%
2ii (Cosmopolitan)	245	242	6816	1	3	98,8%	100,0%	99,9%
2iii (SE Asian-America)	649	649	6413	0	0	100,0%	100,0%	100,0%
2iv (Asian II)	41	33	7021	0	8	80,5%	100,0%	99,9%
2v (Asian I)	317	317	6744	1	0	100,0%	100,0%	100,0%
2vi (Sylvatic)	17	17	7045	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							
Dengue Virus Serotype 3	897	897	6165	0	0	100,0%	100,0%	100,0%
3i	68	68	6992	2	0	100,0%	100,0%	100,0%
3ii	190	189	6872	0	1	99,5%	100,0%	100,0%
3iii	620	620	6442	0	0	100,0%	100,0%	100,0%
3v	18	18	7044	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							
Dengue Virus Serotype 4	216	216	6846	0	0	100,0%	100,0%	100,0%
4i	23	23	7038	1	0	100,0%	100,0%	100,0%
4ii	187	186	6875	0	1	99,5%	100,0%	100,0%
4iii	2	2	7060	0	0	100,0%	100,0%	100,0%
4iv	3	3	7059	0	0	100,0%	100,0%	100,0%
Genotype could not be assigned*	1							

assigned*								
Total DENV	4118							
Chikungunya Virus E1								
East-Central-South-African	1699	1699	5363	0	0	100,0%	100,0%	100,0%
Asian and Caribbean	780	780	6282	0	0	100,0%	100,0%	100,0%
West African	52	52	7010	0	0	100,0%	100,0%	100,0%
Total CHIKV	2531							
Zika Virus ENV								
African	26	26	7036	0	0	100,0%	100,0%	100,0%
Asian	387	387	6675	0	0	100,0%	100,0%	100,0%
Total ZIKV	413							
Total sequences	7062							

S1 File. Accession number of the sequences collected from DENV, ZIKV and CHIKV whole-genome genomes.

A GenBank mining of sequences was performed against whole-genome genomes of these viruses that had the genotype reported for sensitivity, specificity and accuracy tests of the tool.

Download here: <https://doi.org/10.1371/journal.pntd.0007231.s007>

S2 File. Accession number of the sequences collected from DENV, ZIKV and CHIKV envelope genomes.

A GenBank mining of sequences was performed against envelope genomes of these viruses that had the genotype reported for sensitivity, specificity and accuracy tests of the tool.

Download here: <https://doi.org/10.1371/journal.pntd.0007231.s008>

Appendix 3 Supplementary material to the manuscript entitled “Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes”



CORONAVIRUS TYPING TOOL

NO JOBS IN QUEUE

CORONAVIRUS TYPING TOOL DETAILS

Version 1.9

SEQUENCE ASSIGNMENT

Name MN908947.3

Length 29820

VIRUS ASSIGNMENT

Virus assignment Severe acute respiratory syndrome-related coronavirus

CLADE AND GENOTYPE RESULT

Clade assignment SARS-CoV-2 (2019 outbreak)

Supported with phylogenetic analysis and bootstrap 100.0 (≥ 70.0)

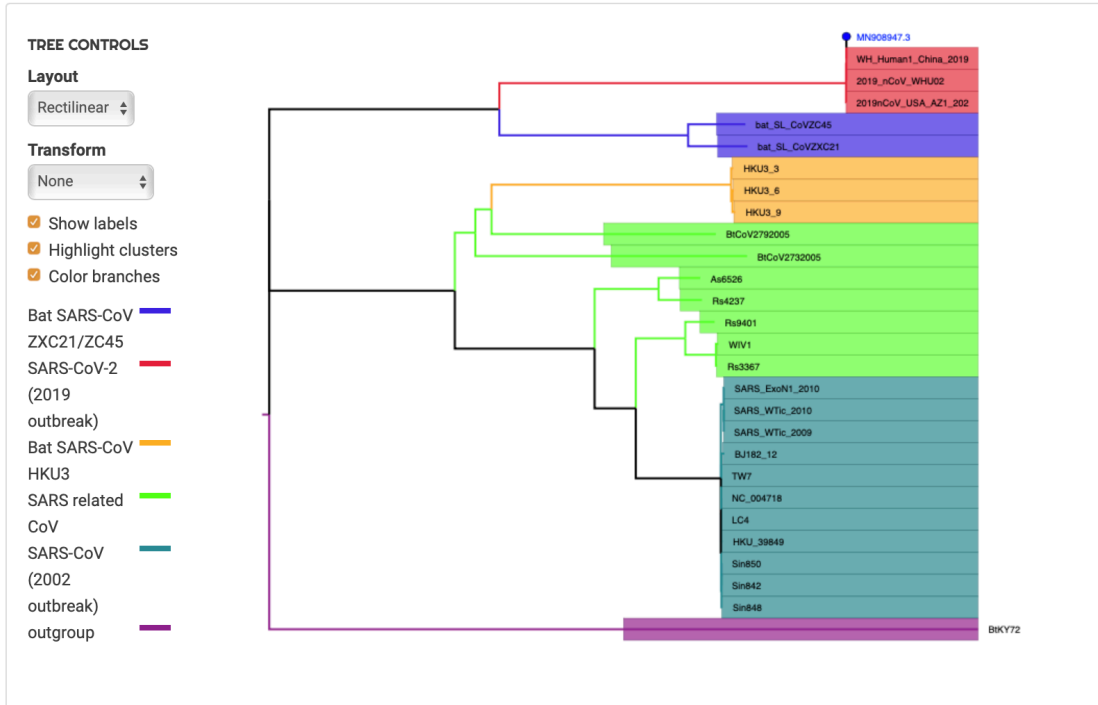
GENOME REGION

Sequence starts at position 1 and ends at position 29820 relative to the NC_045512.2 reference sequence for Wuhan seafood market pneumonia virus (taxon:2697049).

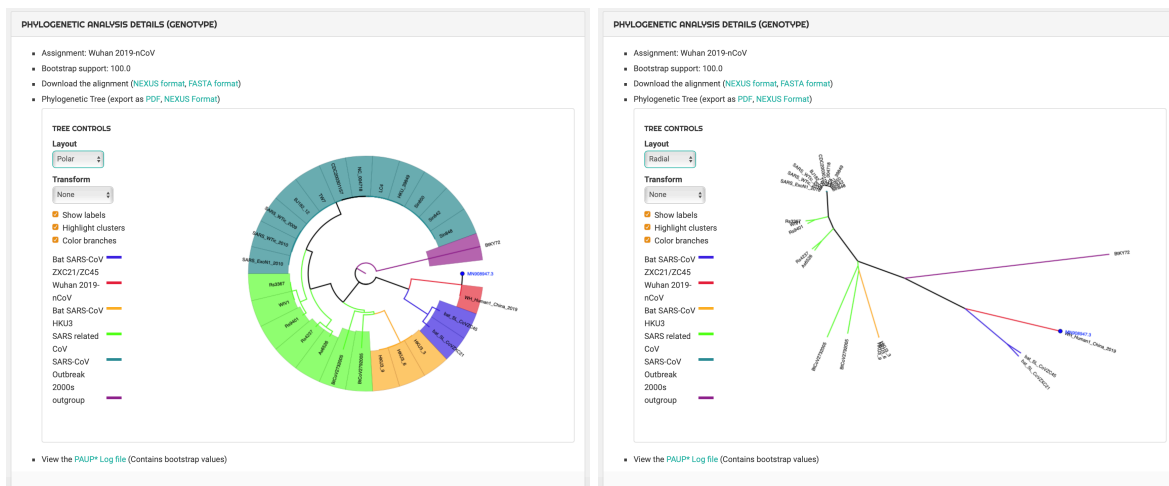


PHYLOGENETIC ANALYSIS DETAILS (GENOTYPE)

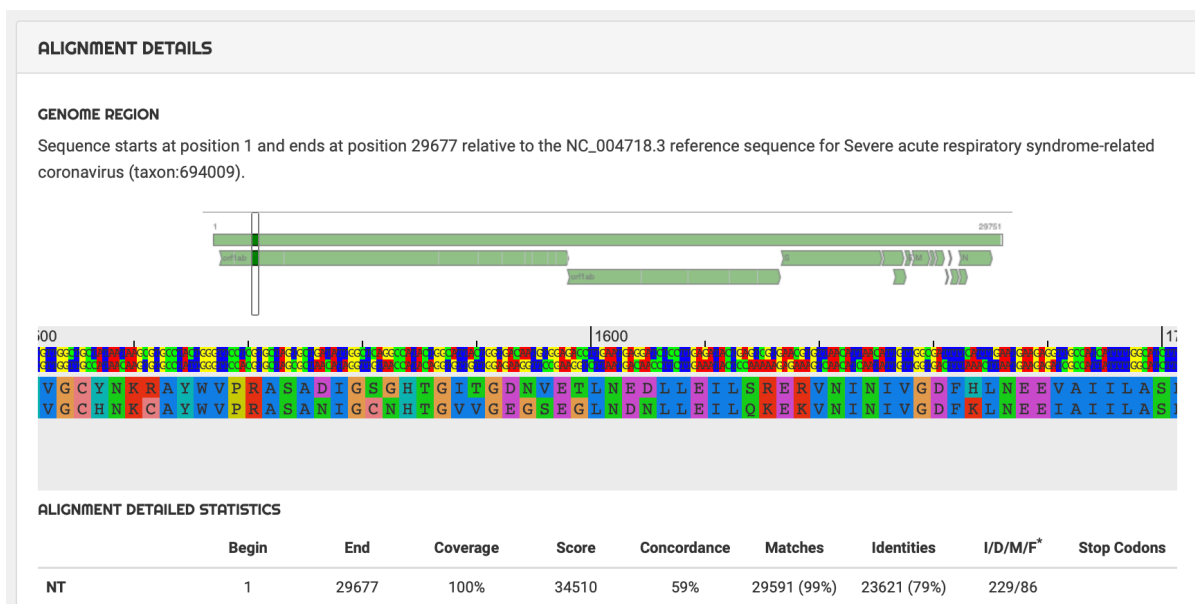
- Assignment: SARS-CoV-2 (2019 outbreak)
- Bootstrap support: 100.0, bootstrap inside 147.1, bootstrap outside 0.0
- Download the alignment ([NEXUS format](#), [FASTA format](#))
- Phylogenetic Tree (export as [PDF](#), [NEXUS Format](#))



Phylogenetic tree can be visualized in polar or radial format:



Alignment Details and Genome Mapper:



Nucleotide (NT) and Coding Regions (CDS) Mutational Analysis

[SHOW MUTATIONS](#)

	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F [±]	Stop Codons
NT	1	29677	99.8%	34510	58.6%	29591 (98.9%)	23621 (79.0%)	229/86	
CDS									
1_orf1ab	1	7074	100%	44034	89.0%	7068 (99.5%)	6126 (86.2%)	29/6/0/0	1
2_orf1ab	1	4383	100%	25297	83.9%	4377 (99.2%)	3553 (80.5%)	29/6/0/0	1
3_S	1	1256	100%	7176	81.6%	1249 (97.5%)	976 (76.2%)	25/7/0/0	1
4_sars3a	1	275	100%	1490	75.5%	275 (99.6%)	200 (72.5%)	1/0/0/0	1
5_sars3b	1	155	100%	519	52.9%	155 (99.4%)	89 (57.1%)	1/0/0/0	5
6_E	1	77	100%	447	96.3%	76 (98.7%)	73 (94.8%)	0/1/0/0	1
7_M	1	222	100%	1419	92.9%	222 (99.6%)	202 (90.6%)	1/0/0/0	1
8_sars6	1	64	100%	300	74.6%	62 (96.9%)	43 (67.2%)	0/2/0/0	1
9_sars7a	1	123	100%	758	89.3%	122 (99.2%)	105 (85.4%)	0/1/0/0	1
10_sars7b	1	45	100%	252	83.7%	44 (97.8%)	36 (80.0%)	0/1/0/0	1
11_sars8a	1	40	100%	100	33.8%	39 (90.7%)	13 (30.2%)	3/1/0/0	0
12_sars8b	1	85	100%	26	4.2%	81 (81.8%)	30 (30.3%)	14/4/1/1	4
13_N	1	423	100%	2645	91.3%	420 (99.3%)	383 (90.5%)	0/3/0/0	1
14_sars9b	1	99	100%	451	72.9%	98 (99.0%)	72 (72.7%)	0/1/0/0	1

Protein Analysis (from UNIPROT RefSeq):

Proteins

orf1ab polyprotein...	1	7074	100%	44034	89.0%	7068 (99.5%)	6126 (86.2%)	29/6/0/0	1
leader protein (NP...	1	180	100%	1059	87.4%	180 (100%)	152 (84.4%)	0/0/0/0	0
counterpart of MH...	1	638	100%	3197	71.2%	638 (100%)	436 (68.3%)	0/0/0/0	0
nsp3-pp1a/pp1ab ...	1	1922	100%	10415	80.4%	1916 (98.2%)	1486 (76.2%)	29/6/0/0	0
nsp4-pp1a/pp1ab ...	1	500	100%	3038	85.4%	500 (100%)	400 (80.0%)	0/0/0/0	0
3C-like proteinase ...	1	306	100%	2172	97.4%	306 (100%)	294 (96.1%)	0/0/0/0	0
nsp6-pp1a/pp1ab ...	1	290	100%	1813	89.9%	290 (100%)	253 (87.2%)	0/0/0/0	0
nsp7-pp1a/pp1ab ...	1	83	100%	508	98.6%	83 (100%)	82 (98.8%)	0/0/0/0	0
nsp8-pp1a/pp1ab ...	1	198	100%	1210	97.6%	198 (100%)	193 (97.5%)	0/0/0/0	0
nsp9-pp1a/pp1ab ...	1	113	100%	752	96.9%	113 (100%)	110 (97.3%)	0/0/0/0	0
formerly known as...	1	139	100%	1061	97.7%	139 (100%)	135 (97.1%)	0/0/0/0	0
RNA-dependent R...	1	932	100%	6561	97.1%	932 (100%)	898 (96.4%)	0/0/0/0	0
nsp13-pp1ab (ZD, ...	1	601	100%	4241	99.9%	601 (100%)	600 (99.8%)	0/0/0/0	0
3'-to-5' exonucleas...	1	527	100%	3864	96.5%	527 (100%)	501 (95.1%)	0/0/0/0	0
endoRNAse (NP_8...	1	346	100%	2174	92.6%	346 (100%)	307 (88.7%)	0/0/0/0	0
2'-O-ribose methylt...	1	298	100%	1968	95.4%	298 (100%)	278 (93.3%)	0/0/0/0	0
orf1a polyprotein (...	1	4383	100%	25297	83.9%	4377 (99.2%)	3553 (80.5%)	29/6/0/0	1
nsp11-pp1a (NP_9...	1	13	100%	71	89.9%	13 (100%)	11 (84.6%)	0/0/0/0	0
E2 glycoprotein pr...	1	1256	100%	7176	81.6%	1249 (97.5%)	976 (76.2%)	25/7/0/0	1

SARS-CoV-2 Protein Mutations and Codon Mutations on E and Matrix Proteins as compared with SARSr-CoV RefSeq (GenBank: NC_004718.3).

protein E (NP_828...	1	77	100%	447	96.3%	76 (98.7%)	73 (94.8%)	0/1/0/0	1
Protein mutations:	T55S (26279A>T 26281G>T), V56F (26282G>T), E69del (26321_26323delGAA), G70R (26324G>A)								
Codon mutations:	GAA8GAG (26140A>G), GTC29GTT (26203C>T), TTA51CTT (26267T>C 26269A>T), CCA54CCT (26278A>T), ACG55TCT (26279A>T 26281G>T), GTT56TTT (26282G>T), GTC58GTT (26290C>T), TCG60TCT (26296G>T), AAC66AAT (26314C>T), GAA69del (26321_26323delGAA), GGA70AGA (26324G>A)								
matrix protein (NP...	1	222	100%	1419	92.9%	222 (99.6%)	202 (90.6%)	1/0/0/0	1
Protein mutations:	D3_N4insS (26405_26406insTTC), Q14K (26437C>A 26439A>G), A29T (26482G>A 26484C>A), M32C (26491A>T 26492T>G 26493G>T), S39A (26512T>G 26514T>C), V51I (26548G>A), V75I (26620G>A 26622G>C), I86L (26653A>C), V96I (26683G>A), R124H (26768G>A 26769G>T), V128L (26779G>C), M133L (26794A>C 26796G>A), I144L (26827A>C), M150I (26847G>T), S154H (26857T>C 26858C>A 26859C>T), G187A (26957G>C 26958C>A), T188G (26959A>G 26960C>G), N196S (26984A>G 26985C>T), A210S (27025G>T), G211S (27028G>A), N213S (27035A>G 27036C>T)								
Codon mutations:	GAC3GAT (26405_26406insTTC), GAC3_AAC4insTCC (26405_26406insTTC), GAG10GAA (26427G>A), CAA14AAG (26437C>A 26439A>G), CTG16CTT (26445G>T), CTA28CTT (26481A>T), GCC29ACA (26482G>A 26484C>A), ATG32TGT (26491A>T 26492T>G 26493G>T), TTA33CTT (26494T>C 26496A>T), TCT39GCC (26512T>G 26514T>C), AAT40AAC (26517T>C), CGG41AGG (26518C>A), AAC42AAT (26523C>T), TAC46TAT (26535C>T), ATA48ATT (26541A>T), CTT50TTA (26545C>T 26547T>A), GTT51ATT (26548G>A), CTC55CTG (26562C>G), TTG56TTA (26565G>A), ACA60ACT (26577A>T), CTT61TTA (26578C>T 26580T>A), GTC69GTT (26604C>T), ATT72ATA (26613T>A), GTG75ATC (26620G>A 26622G>C), ACT76ACC (26625T>C), GGC77GGT (26628C>T), GGG78GGA (26631G>A), GCG80GCT (26637G>T), ATT81ATC (26640T>C), ATT86CTT (26653A>C), CTT92CTC (26673T>C), GTT96ATT (26683G>A), TCC98TCT (26691C>T), AGG100AGA (26697G>A), GCT103GCG (26706T>G), ACC105ACG (26712C>G), CGC106CGT (26715C>T), TCA107TCC (26718A>C), AAC112AAT (26733C>T), ACA115ACT (26742A>T), AAT120AAC (26757T>C), CCT122CCA (26763T>A), GGG124CAT (26768G>A 26769G>T), GGG125GGC (26772G>C), ACA126ACT (26775A>T), GTG128CTG (26779G>C), CTC132CTT (26793C>T), ATG133CTA (26794A>C 26796G>A), CTT137CTC (26808T>C), GTC138GTA (26811C>A), ATT139ATC (26814T>C), GGT140GGA (26817T>A), ATT144CTT (26827A>C), GGT146GGA (26835T>A), CAC147CAT (26838C>T), TTG148CTT (26839T>C 26841G>T), CGA149CGT (26844A>T), ATG150ATT (26847G>T), GCC151GCT (26850C>T), TCC154CAT (26857T>C 26858C>A 26859C>T), GGG156GGA (26865G>A), ATT160ATC (26877T>C), CCA164CCT (26889A>T), GAG166GAA (26895G>A), GTG169GTT (26904G>T), TTA180TTG (26937A>G), GCG182GCT (26943G>T), GGC187GCA (26957G>C 26958C>A), ACT188GGT (26959A>G 26960C>G), GAT189GAC (26964T>C), AAC196AGT (26984A>G 26985C>T), CGT199AGG (26992C>A 26994T>G), GGA201GGC (27000A>C), AAT206AAC (27015T>C), CAC209CAT (27024C>T), GCC210TCC (27025G>T), GGT211AGT (27028G>A), AAC213AGT (27035A>G 27036C>T), CTA219CTT (27054A>T)								

SARS-CoV-2 Protein Mutations and Codon Mutations on E and Matrix Proteins as compared with Bat SARS related CoV sequence, bat_SL_CovZXC21 (GenBank: MG772934)

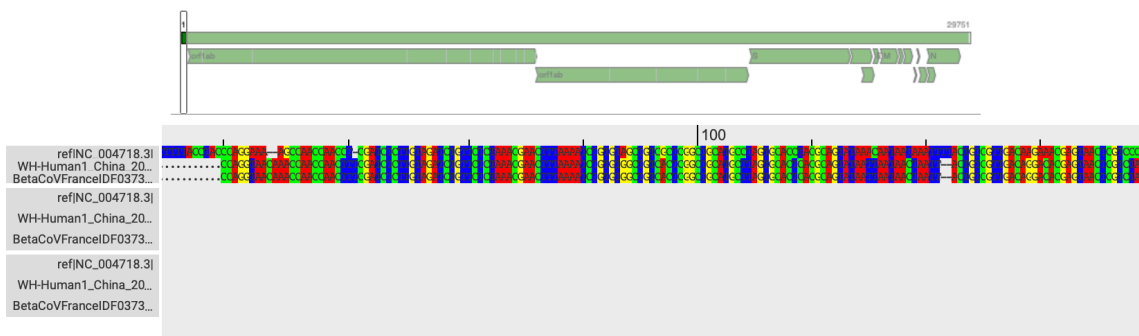
protein E (NP_828...	1	77	98.7%	474	100%	76 (100%)	76 (100%)	0/0/0/0	1
Protein mutations:	none								
Codon mutations:	TTT23TTC (26185T>C), GTC29GTT (26203C>T), TTG75CTG (26399T>C)								
matrix protein (NP...	1	223	100%	1521	98.6%	223 (100%)	220 (98.7%)	0/0/0/0	1
Protein mutations:	S2A (26401T>G), G3D (26405G>A), D311S (26405I2G>T 26405I3A>C)								
Codon mutations:	TCA2GCA (26401T>G), GGT3GAT (26405G>A), GAC31TCC (26405I2G>T 26405I3A>C), ACC6ACT (26415C>T), TTA16CTT (26443T>C 26445A>T), GGA24GGT (26469A>T), TTG26CTA (26473T>C 26475G>A), TTT27TTC (26478T>C), TTG33CTT (26494T>C 26496G>T), TTA34CTA (26497T>C), TAC46TAT (26535C>T), CTT56TTA (26563C>T 26565T>A), TGC63TGT (26586C>T), AAC73AAT (26616C>T), ACT76ACC (26625T>C), GCG80GCT (26637C>T), ATT81ATC (26640T>C), GGC84GCT (26649C>T), CTT92CTC (26673T>C), AGG100AGA (26697G>A), GCT103GCG (26706T>G), TTT111TTC (26730T>C), AAC112AAT (26733C>T), TTG119CTC (26752T>C 26754G>C), CCT122CCA (26763T>A), CTT123CTC (26766T>C), ACA126ACT (26775A>T), AGG130AGA (26787G>A), GAG134GAA (26799G>A), ATT139ATC (26814T>C), GCA151GCT (26850A>T), CTG155CTA (26862G>A), CCC164CCT (26889C>T), GTA169GTT (26904A>T), GGT201GGC (27000T>C), AAT202AAC (27003T>C), TAC203TAT (27006C>T), AAT206AAC (27015T>C)								

SARS-CoV-2 comparison between sequences isolated in France, Jan 2020 (BetaCoV/France/IDF0373/2020) and Wuhan, China, Dec 2019 (GenBank: MN908947)

ALIGNMENT DETAILS

ALIGNMENT

Using **NC_004718.3** (Severe acute respiratory syndrome-related coronavirus (taxon:694009)) as reference for alignment, numbering and genome annotations.



GENETIC DIVERSITY ANALYSIS

Showing genetic similarity and mutations of the sequence against a reference of choice.

Reference for genetic diversity:

[HIDE MUTATIONS](#)

	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F*	Stop Codons
NT	20	29914	99.4%	59610	99.9%	29809 (100%)	29807 (99.9%)	0/0	

Mutations: 22551G>T, 26016G>T

E2 glycoprotein pr...	1	1281	99.5%	8985	99.9%	1274 (100%)	1273 (99.9%)	0/0/0/0	1
Protein mutations: V354F (22551G>T)									
Codon mutations: GTC354TTC (22551G>T)									
hypothetical protei...	1	276	100%	1944	99.4%	276 (100%)	275 (99.6%)	0/0/0/0	1
Protein mutations: G250V (26016G>T)									
Codon mutations: GGT250GTT (26016G>T)									
hypothetical protei...	1	156	100%	964	99.6%	156 (100%)	155 (99.4%)	0/0/0/0	5
Protein mutations: V110F (26016G>T)									
Codon mutations: GTT110TTT (26016G>T)									
protein E (NP_828...	1	77	98.7%	474	100%	76 (100%)	76 (100%)	0/0/0/0	1

Submission page: Input sequences in FASTA format

CORONAVIRUS TYPING TOOL

This tool is designed to use Blast and phylogenetic methods in order to identify the Coronavirus types and genotypes of a nucleotide sequence.

The Coronavirus typing tool also includes a Wuhan Coronavirus genome, this sequence was generously shared by Professor Yong-Zhen Zhang and colleagues via [this post on virological.org](#). Professor Yong-Zhen Zhang and colleagues ask that you communicate with them if you wish to publish results that use the sequence that they share in a journal. We gratefully acknowledge their contribution.

Note for batch analysis: The tool accepts up to 2000 sequences at a time.

INPUT

Submit one or more FASTA sequences to be typed individually. If you have raw NGS reads (short reads or long reads), please use the [Genome Detective Virus Tool](#) to assemble first. Subtyping tools will be linked in the results. [Click here](#) to load some sample data.

Sequence

CLICK OR DROP FILE

```

>MN908947.3 Wuhan seafood market pneumonia
virus isolate Wuhan-Hu-1, complete genome
ATTAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTGAT
CTCTTGTAGATCTGTTCTAAACGAACTTTAAAATCTGTGTGGCTG
TCACTCGGCTGCATGCTTAGTGCCTCACGCAGTATAATTAATAACT
AATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAG
GCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTA
GGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCT
TGGTTTCAACGAGAAAACACACGTCCAACCTCAGTTTGCTGTTTTAC
AGGTTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAG
GTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTT
AGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATG
TGTTTCATCAAACGTTTCGGATGCTCGAAGTGCACCTCATGGTCATGTT
ATGGTTGAGCTGGTAGCAGAAGTTCGAAGGCATTTCAGTACGGTCGTAG
TGGTGAGACACTTGGTGTCTTGTCCCTCATGTGGGCGAAAATACCAG
TGGTTCGCGAAGCTTCTCTGCTTAAACCTTATAAGGACCTT

```

START FREE ANALYSIS

CLEAR

[Log in](#) or [register](#) to experience the advantages of a [premium account](#).

Supplementary Table 1: n2019-CoV GISAID Tested Sequences. GISAID acknowledges the Authors and the Laboratories for their sequence and metadata shared through GISAID. All Originating laboratories are gratefully acknowledged for sharing the data and are cited in this table:

GISAID Accession ID	Virus name	Location	Collection date	Lab citation
EPI_ISL_402125	BetaCoV/Wuhan-Hu-1/2019	China	2019-12	1
EPI_ISL_405839	BetaCoV/Guangdong/20SF013/2031	China / Guangdong / Shenzhen	2020-01	2
EPI_ISL_406030	BetaCoV/Guangdong/20SF013/2032	China / Guangdong / Shenzhen	2020-01	2

EPI_ISL_402119	BetaCoV/Wuhan/IVDC-HB-01/2019	China / Hubei Province / Wuhan City	2019-12-30	3
EPI_ISL_402120	BetaCoV/Wuhan/IVDC-HB-04/2020	China / Hubei Province / Wuhan City	2020-01-01	3
EPI_ISL_402121	BetaCoV/Wuhan/IVDC-HB-05/2019	China / Hubei Province / Wuhan City	2019-12-30	3
EPI_ISL_402123	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	China / Hubei Province / Wuhan City	2019-12-24	4
EPI_ISL_402124	BetaCoV/Wuhan/WIV04/2019	China / Hubei Province / Wuhan City	2019-12-30	5
EPI_ISL_402127	BetaCoV/Wuhan/WIV02/2019	China / Hubei Province / Wuhan City	2019-12-30	5
EPI_ISL_402128	BetaCoV/Wuhan/WIV05/2019	China / Hubei Province / Wuhan City	2019-12-30	5
EPI_ISL_402129	BetaCoV/Wuhan/WIV06/2019	China / Hubei Province / Wuhan City	2019-12-30	5
EPI_ISL_402130	BetaCoV/Wuhan/WIV07/2019	China / Hubei Province / Wuhan City	2019-12-30	5
EPI_ISL_403928	BetaCoV/Wuhan/IPBCAMS-WH-05/2020	China / Hubei Province / Wuhan City	2020-01-01	4
EPI_ISL_403929	BetaCoV/Wuhan/IPBCAMS-WH-04/2019	China / Hubei Province / Wuhan City	2019-12-30	4
EPI_ISL_403930	BetaCoV/Wuhan/IPBCAMS-WH-03/2019	China / Hubei Province / Wuhan City	2019-12-30	4
EPI_ISL_403931	BetaCoV/Wuhan/IPBCAMS-WH-02/2019	China / Hubei Province / Wuhan City	2019-12-30	4
EPI_ISL_402131	BetaCoV/bat/Yunnan/RaTG13/2013	China / Yunnan Province / Pu'er City	2013-07-24	6
EPI_ISL_406531	BetaCoV/Guangdong/20SF013/2037	China/Guangdong Province	2020-01-22	7
EPI_ISL_406534	BetaCoV/Guangdong/20SF013/2039	China/Guangdong Province	2020-01-22	7
EPI_ISL_406535	BetaCoV/Guangdong/20SF013/2040	China/Guangdong Province	2020-01-22	7
EPI_ISL_406536	BetaCoV/Guangdong/20SF013/2041	China/Guangdong Province	2020-01-22	7
EPI_ISL_406538	BetaCoV/Guangdong/20SF013/2042	China/Guangdong Province	2020-01-23	7
EPI_ISL_403932	BetaCoV/Guangdong/20SF012/2020	China/Guangdong, China	2020-01-14	7
EPI_ISL_403933	BetaCoV/Guangdong/20SF013/2020	China/Guangdong, China	2020-01-15	7
EPI_ISL_403934	BetaCoV/Guangdong/20SF013/2021	China/Guangdong, China	2020-01-15	7
EPI_ISL_403935	BetaCoV/Guangdong/20SF013/2022	China/Guangdong, China	2020-01-15	7
EPI_ISL_403936	BetaCoV/Guangdong/20SF013/2023	China/Guangdong, China	2020-01-17	7
EPI_ISL_403937	BetaCoV/Guangdong/20SF013/2024	China/Guangdong, China	2020-01-18	7
EPI_ISL_406533	BetaCoV/Guangdong/20SF013/2038	China/Guangzhou City	2020-01-22	7
EPI_ISL_402132	BetaCoV/Wuhan/HBCDC-HB-01/2019	China/Hubei Province	2019-12-30	5
EPI_ISL_406592	BetaCoV/Guangdong/20SF013/2043	China/Shenzhen	2020-01-13	8
EPI_ISL_406593	BetaCoV/Guangdong/20SF013/2044	China/Shenzhen	2020-01-13	9
EPI_ISL_406594	BetaCoV/Guangdong/20SF013/2045	China/Shenzhen	2020-01-16	9
EPI_ISL_406595	BetaCoV/Guangdong/20SF013/2046	China/Shenzhen	2020-01-16	9
EPI_ISL_404227	BetaCoV/Guangdong/20SF013/2027	China/Zhejiang, China	2020-01-16	10
EPI_ISL_404228	BetaCoV/Guangdong/20SF013/2028	China/Zhejiang, China	2020-01-17	10
EPI_ISL_406596	BetaCoV/France/IDF0372/2020 EPI_ISL_406596	France / Ile-de-France / Paris	2020-01-23	11
EPI_ISL_406597	BetaCoV/France/IDF0373/2020 EPI_ISL_406597	France / Ile-de-France / Paris	2020-01-23	11
EPI_ISL_402126	BetaCoV/Kanagawa/1/2020	Japan/ Kanagawa Prefecture, Japan	2020-01-14	12
EPI_ISL_406031	BetaCoV/Guangdong/20SF013/2033	Taiwan/Kaohsiung City	2020-01-23	13
EPI_ISL_403962	BetaCoV/Guangdong/20SF013/2025	Thailand/ Nonthaburi Province	2020-01-08	14
EPI_ISL_403963	BetaCoV/Guangdong/20SF013/2026	Thailand/ Nonthaburi Province	2020-01-13	14

EPI_ISL_406223	BetaCoV/Guangdong/20SF013/2036	USA / Arizona / Phoenix	2020-01-22	15
EPI_ISL_406034	BetaCoV/Guangdong/20SF013/2034	USA / California / Los Angeles	2020-01-23	16
EPI_ISL_406036	BetaCoV/Guangdong/20SF013/2035	USA / California / Orange County	2020-01-22	16
EPI_ISL_404253	BetaCoV/Guangdong/20SF013/2029	USA / Illinois /Chicago	2020-01-21	17
EPI_ISL_404895	BetaCoV/Guangdong/20SF013/2030	USA / Washington / Snohomish County	2020-01-19	18

Citations and acknowledgements to original laboratories generating the data:

- 1 - National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)
- 2 - The University of Hong Kong - Shenzhen Hospital
- 3 - National Institute for Viral Disease Control and Prevention, China CDC
- 4 - Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
- 5 - Wuhan Jinyintan Hospital, Wuhan Institute of Virology, Chinese Academy of Sciences
- 6 - Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College
- 7 - Guangdong Provincial Center for Diseases Control and Prevention; Guangdong Provincial Public Health
- 8 - Wuhan Jinyintan Hospital, Hubei Provincial Center for Disease Control and Prevention
- 9 - Shenzhen Key Laboratory of Pathogen and Immunity, National Clinical Research Center for Infectious Disease, Shenzhen Third People's Hospital
- 10 -Zhejiang Provincial Center for Disease Control and Prevention, Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention
- 11 - Department of Infectious and Tropical Diseases, Bichat Claude Bernard Hospital, Paris, National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris
- 12 -Department of Virology III, National Institute of Infectious Diseases
- 13 - Centers for Disease Control, R.O.C. (Taiwan)
- 14 - Bamrasnaradura Hospital, Department of Medical Sciences, Ministry of Public Health, Thailand, Thai Red Cross Emerging Infectious Diseases - Health Science Centre, Department of Disease Control, Ministry of Public Health, Thailand
- 15 - Arizona Department of Health Services, Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention
- 16 - California Department of Public Health, Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Diseases Control and Prevention
- 17 - Department of Public Health Chicago Laboratory, Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for sControl and Prevention
- 18 - Providence Regional Medical Center, Division of Viral Diseases, Centers for Disease Control and Prevention

Supplementary Table 2: Reference phylogenetic dataset sequences selected for the Genome Detective Coronavirus Tool.

SARSr-CoV Cluster	Sequence Name	Accession Number	Location	Host
Bat SARS-CoV HKU3	HKU3_3	DQ084200	China	Bat
Bat SARS-CoV HKU3	HKU3_6	GQ153541	China	Bat
Bat SARS-CoV HKU3	HKU3_9	GQ153544	China	Bat
Bat SARS-CoV ZXC21/ZC45	bat_SL_CoVZXC21	MG772934	China	Bat
Bat SARS-CoV ZXC21/ZC45	bat-SL-CoVZC45	MG772933.1	China	Bat
SARS related CoV	As6526	KY417142	China	Bat
SARS related CoV	BtCoV2732005	DQ648856	China	Bat
SARS related CoV	BtCoV2792005	DQ648857	China	Bat
SARS related CoV	Rs3367	KC881006	China	Bat
SARS related CoV	Rs4237	KY417147	China	Bat
SARS related CoV	Rs9401	KY417152	China	Bat
SARS related CoV	WIV1	KF367457	China	Bat
SARS-CoV Outbreak 2002-3	BJ182_12	EU371564	China	Human
SARS-CoV Outbreak 2002-3	HKU_39849	AY278491	Hong Kong	Human
SARS-CoV Outbreak 2002-3	LC4	AY395001	China	Human
SARS-CoV Outbreak 2002-3	NC_004718	NC_004718	Canada	Human
SARS-CoV Outbreak 2002-3	SARS_ExoN1_2010	KF514393	USA	N/A
SARS-CoV Outbreak 2002-3	SARS_WTic_2009	KF514394	USA	N/A
SARS-CoV Outbreak 2002-3	SARS_WTic_2010	KF514388	USA	N/A
SARS-CoV Outbreak 2002-3	Sin842	AY559081	Singapore	Human
SARS-CoV Outbreak 2002-3	Sin848	AY559085	Singapore	Human
SARS-CoV Outbreak 2002-3	Sin850	AY559096	Singapore	Human
SARS-CoV Outbreak 2002-3	TW7	AY502930	Taiwan	Human
Wuhan 2019-nCoV	WH-Human1_China_2019-Dec	MN908947	China	Human
2019nCoV_USA_AZ1_2020	2019nCoV_USA_AZ1_2020	MN997409	USA	Human
2019_nCoV_WHU02	2019_nCoV_WHU02	MT019532	China	Human

Supplementary Table 3: Evaluation of the Genome Detective Coronavirus Typing Tool to classify coronavirus complete genomes. The classification results were compared to manual phylogenetic analysis. In this table, the following abbreviations are used: TP, total positives; TN, total negatives; FP, false positive; FN, False negative; Sens, sensitivity; Spec, specificity; PPV, positive predicted value; NPVs negative predicted value, ACC, accuracy.

Virus species	Known	TP	TN	FP	FN	SENS	SPEC	ACC
Betacoronavirus	121	121	311	0	0	100%	100%	100%
Human Coronavirus HKU1	19	19	413	0	0	100%	100%	100%
Longquan RI rat coronavirus	1	0	432	0	1	0%	100%	100%
MERSr-CoV	97	97	335	0	0	100%	100%	100%
Murine Hepatitis Virus	9	9	423	0	0	100%	100%	100%

Rat Coronavirus	3	3	429	0	0	100%	100%	100%
Rousettus bat coronavirus HKU9	4	4	428	0	0	100%	100%	100%
SARSr-CoV	176	176	256	0	0	100%	100%	100%
Tylonycteris bat coronavirus HKU4	1	1	431	0	0	100%	100%	100%
Zaria_bat_coronavirus	1	0	432	0	1	0%	100%	100%
Total	432							
SARSr-CoV Clusters	Known	TP	TN	FP	FN	SENS	SPEC	ACC
<i>Bat SARS-CoV HKU3</i>	8	8	424	0	0	100%	100%	100%
<i>Bat SARS-CoV ZXC21/ZC45</i>	2	2	430	0	0	100%	100%	100%
<i>SARS related CoV</i>	6	6	426	0	0	100%	100%	100%
<i>SARSr-CoV outbreak 2002-3</i>	112	112	320	0	0	100%	100%	100%
Wuhan 2019-nCoV	47	26	406	0	0	100%	100%	100%
Total	176							
Total sequences	432							

Supplementary Table 4A: Nucleotide and protein mutational analysis performed by Genome Detective Coronavirus Typing Tool. This table compares an n2019-CoV query sequence (GenBank Accession Number MN90847) to the reference strain of SARS (NC_004718.3: GenBank). In the table, the following abbreviations are used: Begin, first nucleotide position that reference sequence; End, last nucleotide position that matches the reference sequence; Coverage: coverage of reference sequence genome; Score, nucleotide and amino acid score of AGA; Matches, number of matches; Identities, number of identical nucleotides; I, insertions; D, deletions; M, misaligned; F, Frameshifts; Stop codons, number of stop codons and CDS, coding sequencing.

Ref: SARS: NC_004718.3	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D	
Query: n2019-CoV: MN90847	1	2967	99.8%	3451	58.6%	29591 (98.9%)	23621 (79.0%)	229/86	
CDS	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F	Stop Codons
1_orf1ab	1	7074	100%	4403	89.0%	7068 (99.5%)	6126 (86.2%)	29/6/0/0	1
2_orf1ab	1	4383	100%	2529	83.9%	4377 (99.2%)	3553 (80.5%)	29/6/0/0	1
3_S	1	1256	100%	7176	81.6%	1249 (97.5%)	976 (76.2%)	25/7/0/0	1
4_sars3a	1	275	100%	1490	75.5%	275 (99.6%)	200 (72.5%)	1/0/0/0	1
5_sars3b	1	155	100%	519	52.9%	155 (99.4%)	89 (57.1%)	1/0/0/0	5
6_E	1	77	100%	447	96.3%	76 (98.7%)	73 (94.8%)	0/1/0/0	1
7_M	1	222	100%	1419	92.9%	222 (99.6%)	202 (90.6%)	1/0/0/0	1
8_sars6	1	64	100%	300	74.6%	62 (96.9%)	43 (67.2%)	0/2/0/0	1
9_sars7a	1	123	100%	758	89.3%	122 (99.2%)	105 (85.4%)	0/1/0/0	1
10_sars7b	1	45	100%	252	83.7%	44 (97.8%)	36 (80.0%)	0/1/0/0	1
11_sars8a	1	40	100%	100	33.8%	39 (90.7%)	13 (30.2%)	3/1/0/0	0
12_sars8b	1	85	100%	26	4.2%	81 (81.8%)	30 (30.3%)	14/4/1/1	4

13_N	1	423	100%	2645	91.3%	420 (99.3%)	383 (90.5%)	0/3/0/0	1
14_sars9b	1	99	100%	451	72.9%	98 (99.0%)	72 (72.7%)	0/1/0/0	1

Supplementary Table 4B: Nucleotide and protein mutational analysis performed by Genome Detective Coronavirus Typing Tool. This table compares an n2019-CoV query sequence (GenBank Accession Number MN90847) to the Bat SARS related CoV sequence, bat_SL_CovZXC21 (MG772934: GenBank). In the table, the following abbreviations are used: Begin, first nucleotide position that reference sequence; End, last nucleotide position that matches the reference sequence; Coverage: coverage of reference sequence genome; Score, nucleotide and amino acid score of AGA; Matches, number of matches; Identities, number of identical nucleotides; I, insertions; D, deletions; M, misaligned; F, Frameshifts; Stop codons, number of stop codons and CDS, coding sequencing.

B) bat_SL_CoVZXC21: MG772934	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D	
n2019-CoV: MN908947	1	2974	99.5%	4475	75.6%	29638 (99.3%)	26108 (87.5%)	182/17	
CDS	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F	Stop Codons
1_orf1ab	1	7078	99.9%	4759	96.4%	7071 (99.6%)	6767 (95.3%)	28/1/6/ 0	1
2_orf1ab	1	4387	99.9%	2893	96.3%	4380 (99.3%)	4197 (95.2%)	28/1/6/ 0	1
3_S	1	1253	99.4%	7104	83.8%	1243 (97.6%)	1015 (79.7%)	29/2/10 /0	1
4_sars3a	1	276	100%	1847	93.7%	276 (100%)	254 (92.0%)	0/0/0/0	1
5_sars3b	1	156	100%	755	75.1%	156 (100%)	124 (79.5%)	0/0/0/0	5
6_E	1	77	98.7%	474	100%	76 (100%)	76 (100%)	0/0/0/0	1
7_M	1	223	100%	1521	98.6%	223 (100%)	220 (98.7%)	0/0/0/0	1
8_sars6	1	64	96.9%	369	86.6%	62 (100%)	58 (93.5%)	0/0/0/0	1
9_sars7a	1	123	99.2%	776	91.0%	122 (100%)	108 (88.5%)	0/0/0/0	1
10_sars7b	1	45	97.8%	288	93.8%	44 (100%)	41 (93.2%)	0/0/0/0	1
11_sars8a	1	43	97.7%	331	98.5%	42 (100%)	40 (95.2%)	0/0/0/0	0
12_sars8b	1	99	96.0%	453	74.0%	95 (100%)	73 (76.8%)	0/0/2/0	4
13_N	1	422	99.5%	2707	94.6%	419 (99.5%)	396 (94.1%)	1/1/0/0	1
14_sars9b	1	99	99.0%	446	72.3%	98 (100%)	72 (73.5%)	0/0/0/0	1

Supplementary Table 4C: Nucleotide and protein mutational analysis performed by Genome Detective Coronavirus Typing Tool. This table compares an n2019-CoV query sequence, BetaCoV/France/IDF0373/2020|EPI_ISL_406597, isolated in France, Jan 2010, (GISAID Accession Number ISL_406597) to the first n2019-CoV sequence, isolated in Wuhan, Dec 2019 (MN908947: GenBank). The following abbreviations are used: Begin, first nucleotide position that reference sequence; End, last nucleotide position that matches the reference sequence; Coverage: coverage of reference sequence genome; Score, nucleotide and amino acid score of AGA; Matches, number of matches; Identities, number of identical nucleotides; I, insertions; D, deletions; M, misaligned; F, Frameshifts; Stop codons, number of stop codons and CDS, coding sequencing.

C) n2019-CoV: MN908947	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D	
n2019-CoV: BetaCoV/France/IDF0373/2020	20	29914	99.4%	59606	99.9%	29809 (100%)	29806 (99.9%)	0/0	
CDS	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F	Stop Codons
1_orf1ab	1	7103	99.9%	49596	100%	7097 (100%)	7097 (100%)	0/0/0/0	1
2_orf1ab	1	4412	99.9%	30277	100%	4406 (100%)	4406 (100%)	0/0/0/0	1
3_S	1	1281	99.5%	8985	99.9%	1274 (100%)	1273 (99.9%)	0/0/0/0	1
4_sars3a	1	276	100%	1944	99.4%	276 (100%)	275 (99.6%)	0/0/0/0	1
5_sars3b	1	156	100%	964	99.6%	156 (100%)	155 (99.4%)	0/0/0/0	5
6_E	1	77	98.7%	474	100%	76 (100%)	76 (100%)	0/0/0/0	1
7_M	1	223	100%	1538	100%	223 (100%)	223 (100%)	0/0/0/0	1
8_sars6	1	64	96.9%	400	100%	62 (100%)	62 (100%)	0/0/0/0	1
9_sars7a	1	123	99.2%	846	100%	122 (100%)	122 (100%)	0/0/0/0	1
10_sars7b	1	45	97.8%	316	100%	44 (100%)	44 (100%)	0/0/0/0	1
11_sars8a	1	43	97.7%	339	100%	42 (100%)	42 (100%)	0/0/0/0	0
12_sars8b	1	99	96.0%	608	100%	95 (100%)	95 (100%)	0/0/2/0	4
13_N	1	423	99.3%	2860	99.9%	420 (100%)	419 (99.8%)	0/0/0/0	1
14_sars9b	1	99	99.0%	618	100%	98 (100%)	98 (100%)	0/0/0/0	1

Supplementary Table 4C-II: This table compares an n2019-CoV query sequence, BetaCoV/France/IDF0373/2020|EPI_ISL_406597, isolated in France, Jan 2010, (GISAID Accession Number ISL_406597) to the first n2019-CoV sequence, isolated in Wuhan, Dec 2019 (MN908947: GenBank). The French isolated sequence has two nucleotide (NT) mutations at genome position 22551 (G to T) and 26016 (G to T). These mutations affect three proteins: E2 glycoprotein (Uniprot Accession Number: NP_828851.1), Hypothetical protein sars3a (Uniprot Accession Number: NP_828852.2) and Hypothetical protein sars3b ((Uniprot Accession Number: NP_828853.1). Please note that NT mutation 26016G>T affects two open reading frames that code two hypothetical proteins, sars3a and sars3b.

Mutations NT	22551G>T, 26016G>T
Proteins	Mutations:
E2 glyco (NP_828851.1)	Protein mutations: V354F (22551G>T) Codon mutations: GTC354TTC (22551G>T)
sars3a (NP_828852.2)	Protein mutations: G250V (26016G>T) Codon mutations:

	GGT250GTT (26016G>T)
sars3b (NP_828853.1)	Protein mutations: V110F (26016G>T) Codon mutations: GTT110TTT (26016G>T)

Appendix 4 Supplementary material to the manuscript entitled “West Nile Virus in Brazil”

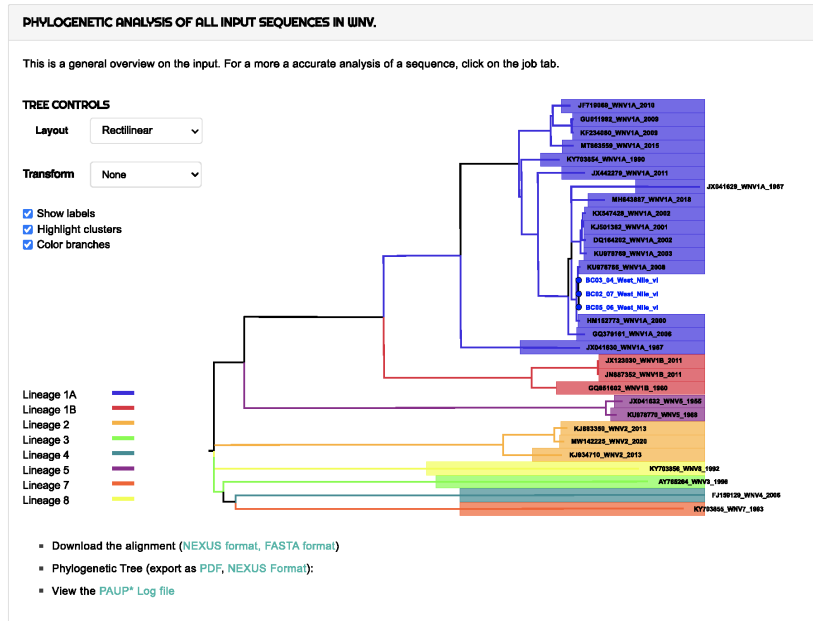


Figure S1. WNV typing tool.

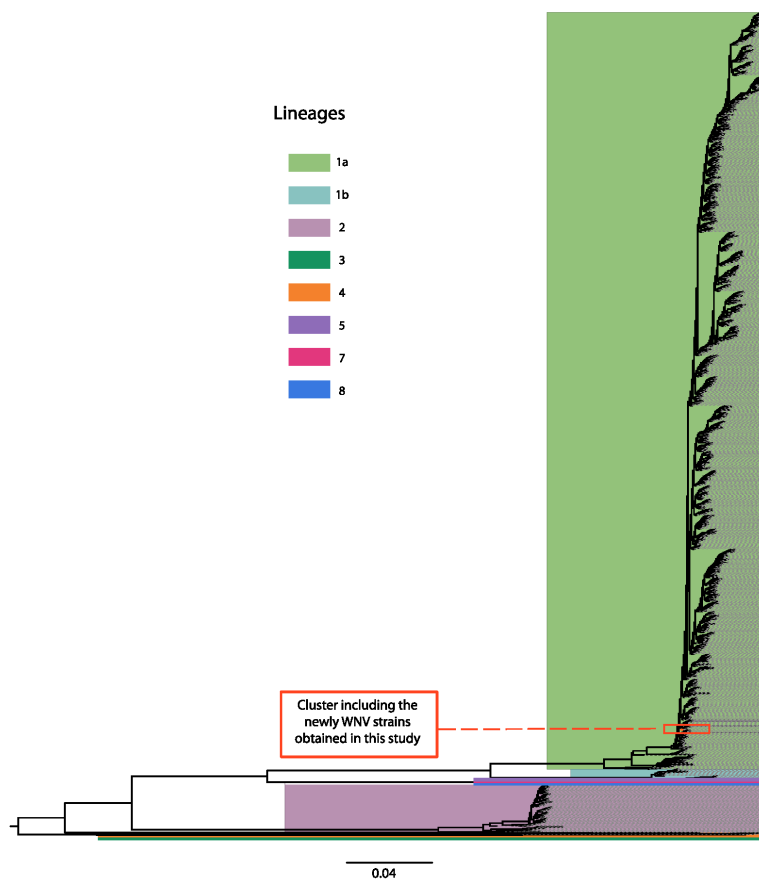


Figure S2. Maximum likelihood phylogenetic tree of 2321 WNV complete genomes. Colours indicates different lineages. Highlighted red clade include the WNV viral strain obtained in this study.

Table S1. Primer scheme.

Primer Name	Sequence
WNVL1a_1_LEFT	GCCTGTGTGAGCTGACAACTT
WNVL1a_1_RIGHT	TTTGTTTTGAGCTCCGCCGATT
WNVL1a_2_LEFT	GGATCGGTGGAGAGGTGTAAT
WNVL1a_2_RIGHT	GACTTTGTGCACCAACAGTCGA
WNVL1a_3_LEFT	TGTCAGAGCAATGGATGTGGGA
WNVL1a_3_RIGHT	CTGTTGCTCATTCCAAGGCAGT
WNVL1a_4_LEFT	GTCATTGGTTGGATGCTTGGGA
WNVL1a_4_RIGHT	TGTGTCAATGCTTCCTTTGCCA
WNVL1a_5_LEFT	AATGACAAAACGTGCTGACCCAG
WNVL1a_5_RIGHT	CACTCACGATGGACCAAGAACG
WNVL1a_6_LEFT	GGAGAATATGGAGAAGTGACAGTGG
WNVL1a_6_RIGHT	AAAGCCTTTGAACAGACGCCAT
WNVL1a_7_LEFT	TTCAAGCAACACTGTCAAGTTGAC
WNVL1a_7_RIGHT	TTTCCAATGCTGCTTCCAGAC
WNVL1a_8_LEFT	CCTGATTGAATTGGAACACCCT
WNVL1a_8_RIGHT	ACGGAGAGGAAGAGCAGAACTC
WNVL1a_9_LEFT	GCATGTCCTGGATAACGCAAGG
WNVL1a_9_RIGHT	AACCACGACACTAAGGTCCACA
WNVL1a_10_LEFT	TCTACGATCAGTTCCAGACTGGA
WNVL1a_10_RIGHT	GTTGTTCTTGACAGCCGTTCCA
WNVL1a_11_LEFT	AGTGGAGGATTTGGATTGGTCT
WNVL1a_11_RIGHT	AGTCAATCTCTACCCGGCCTTC
WNVL1a_12_LEFT	GGCGATGGAATCCTTGAGAGTG
WNVL1a_12_RIGHT	GGCCCACTGAAAAGGGTCAAT
WNVL1a_13_LEFT	CGGCTGTGGTATGGTATGGAG
WNVL1a_13_RIGHT	GAAAACAGCCGCAACATCAAC
WNVL1a_14_LEFT	GGCGACCTCAAGATACAACCA
WNVL1a_14_RIGHT	GCTAGAGCCAAGCATAGCAGAC
WNVL1a_15_LEFT	GGATACTGCTGTGATGGTCGG
WNVL1a_15_RIGHT	TCATCAAGCCGCACATCAACTC
WNVL1a_16_LEFT	TTCTGGGAAATCAACAGATATGTGA
WNVL1a_16_RIGHT	TCAAAGCGGCTCCTTTTGTGT
WNVL1a_17_LEFT	TCTACAGGATCATGACTCGCGG
WNVL1a_17_RIGHT	CGCTTATGTATGAGCCGTGGG
WNVL1a_18_LEFT	GACTTTGGACTTCCCACTGGA
WNVL1a_18_RIGHT	ATTATGTTCTCTGGGCACTGCG
WNVL1a_19_LEFT	ACAGAAGACTGAGAACAGCCGT
WNVL1a_19_RIGHT	TCCAGAGTCCAAGCTCGATCC
WNVL1a_20_LEFT	TATCATGACAGCCACCCACC
WNVL1a_20_RIGHT	GCTGCTACTGCAGATGGTTCTC

WNVL1a_21_LEFT	CTAACTCAAGGCGAGCAGGGT
WNVL1a_21_RIGHT	TGCAGTCCTCAACAGTTCAGA
WNVL1a_22_LEFT	TCTACCAACCAGAGCGTGAGAA
WNVL1a_22_RIGHT	CCCAGAACCTCAATGAGCCCTA
WNVL1a_23_LEFT	GAAAGGAAGATTCTGAGGCCGC
WNVL1a_23_RIGHT	CGTTCCGGAACTTCAGCCATC
WNVL1a_24_LEFT	GTATTCTTCTCCTCATGCAGCG
WNVL1a_24_RIGHT	CGGCCTCAAGTCCAGAAGAAAC
WNVL1a_25_LEFT	GTTGGCTGGACAAGACCAAGAG
WNVL1a_25_RIGHT	GGAACCATGTAGGCATAGTGGC
WNVL1a_26_LEFT	CTTCGTCGATGTTGGAGTGTCG
WNVL1a_26_RIGHT	CTCCATTCTCCAAAGCGTCAC
WNVL1a_27_LEFT	GCTGATCTTAGTGTCTAGCTGC
WNVL1a_27_RIGHT	CAGTTTGCTGTGCCCCTAGAG
WNVL1a_28_LEFT	GTACCGCAAAGAGGCCATCATC
WNVL1a_28_RIGHT	TTGACGAGGACTCTCCGATGTC
WNVL1a_29_LEFT	TGGAACATTGTCACCATGAAGAGT
WNVL1a_29_RIGHT	CTTCTCGTATTGGGGTCCCTT
WNVL1a_30_LEFT	GAGTCGAGCTTCAGGCAATGTG
WNVL1a_30_RIGHT	AGGGAGTAGTGCAGTCATGGC
WNVL1a_31_LEFT	ACGGCAGTTATGATGTGAAGCC
WNVL1a_31_RIGHT	CTCTCATCCACCATCTCCCAA
WNVL1a_32_LEFT	GTCAACAGCAATGCAGCTTGG
WNVL1a_32_RIGHT	GCCAACTCACGCAGGATGTAA
WNVL1a_33_LEFT	TCGAGGCTCTGGGTTTTCTCAA
WNVL1a_33_RIGHT	TTCCCCTCCATCATCCTCACC
WNVL1a_34_LEFT	TCCAGAGAAGATCAGAGGGGGA
WNVL1a_34_RIGHT	TGGAACCACCAAGTGTCTTCCA
WNVL1a_35_LEFT	AGAGTGGAAACCGTCAACTGGA
WNVL1a_35_RIGHT	CCAGACCTCCAACATGCCTCT
WNVL1a_36_LEFT	GTGGCTGCTTCTGTACTCCAC
WNVL1a_36_RIGHT	TCTACAGTACTGTGTCCTCAACCA
WNVL1a_37_LEFT	GTGGCTATCAACCAAGTCAGAGC
WNVL1a_37_RIGHT	CAACATGTGGGGTCTTCTTCC
WNVL1a_38_LEFT	GAAGTTGAGTAGACGGTGCTGC
WNVL1a_38_RIGHT	ACGGGGTCTCCACTAACCTCTA

Table S2. WNV suspected cases reported between 2014-2020 in each Brazilian state, according to SINAN.

States	2014	2015	2016	2017	2018	2019	2020
AC	0	0	0	2	0	0	0
AL	2	0	0	0	0	0	0
AM	1	2	0	0	0	0	0
AP	0	0	0	0	0	0	0
BA	0	0	0	0	0	0	0
CE	0	0	0	0	0	1	1
DF	0	0	0	1	0	0	0
ES	0	0	0	1	49	13	0
GO	0	0	0	0	1	2	0
MA	0	1	0	0	1	1	0
MG	2	0	0	0	1	2	0
MS	0	0	0	0	0	0	0
MT	0	0	0	0	2	0	0
PA	0	1	0	0	1	0	0
PB	0	0	0	0	0	0	0
PE	0	0	0	0	0	2	0
PI	10	118	76	2	3	18	5
PR	0	1	0	1	0	1	0
RJ	0	0	0	0	0	0	0
RN	1	0	0	1	4	1	0
RO	0	0	0	0	0	0	0
RR	0	1	1	0	2	0	0
RS	0	0	2	0	0	0	0
SC	0	0	0	1	0	0	0
SE	0	0	0	0	0	0	0
SP	2	5	0	3	8	6	0
TO	0	0	0	0	2	0	0

Table S3. Globally reference WNV sequences from the subset n=29 used in this study.

Accession Number	Collection date	Country	Lineage	Host
MN849176	01/09/2019	USA	1a	Mosquito
JF719069	01/08/2010	Spain	1a	Horse
KY703854	01/01/1990	Senegal	1a	Mosquito
GU011992	01/01/2009	Italy	1a	Human
KF234080	01/01/2009	Italy	1a	Human
MT863559	10/03/2015	France	1a	Horse
JX442279	01/01/2011	China	1a	Mosquito
JX041629	01/01/1967	Azerbaijan	1a	Bird
JX041630	01/01/1967	Azerbaijan	1a	Bird
KX547428	26/09/2002	USA	1a	Mosquito
KJ501362	01/01/2001	USA	1a	Crow
DQ164202	01/01/2002	USA	1a	Human
MH643887	26/04/2018	Brazil	1a	Horse
KU978769	30/05/2003	Mexico	1a	Crow
KU978766	01/07/2009	Colombia	1a	Flamingo
GQ379161	01/02/2006	Argentina	1a	Horse
HM152773	01/01/2000	Israel	1a	Human
JX123030	03/07/2011	Australia	1b	Horse
GQ851602	01/01/1960	Australia	1b	Mosquito
JN887352	01/01/2011	Australia	1b	Horse
KJ883350	05/07/1905	Greece	2	Human
KJ934710	27/08/2013	Romania	2	Tick
MW142225	01/08/2020	Germany	2	Human
AY765264	01/01/1998	Czech Republic	3	Mosquito
FJ159129	01/01/2006	Russia	4	Mosquito
JX041632	01/01/1955	india	5	Mosquito
KU978770	02/12/1988	india	5	Human
KY703855	01/01/1993	Senegal	7	Tick
KY703856	01/01/1992	Senegal	8	Mosquito