The evolution, molecular epidemiology and genomic sampling of emerging viruses

Bernardo Miguel Gutierrez Granja

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2022

The evolution, molecular epidemiology and genomic sampling of emerging viruses

Bernardo Miguel Gutierrez Granja, Pembroke College, University of Oxford

A thesis submitted for the degree of Doctor of Philosophy, Hilary, 2022

Abstract

Emerging viruses represent a challenging threat to global health. Increasing human travel and global connectivity, changing climate patterns and incremental contact rates between human and animal populations contribute to the likelihood of emerging viral epidemics to occur at greater scales and higher frequencies. These pathogens have gained increased notoriety in recent years as emerging and re-emerging viruses such as Ebola virus, Zika Virus and SARS-CoV-2 caused large epidemics and pandemics of high consequence. The circulation and evolution of these viruses in regions like Africa and the Americas (where genomic surveillance remains limited) is key to their transmission dynamics, driven by environmental conditions combined with abundant biodiversity and complex human-animal interactions.

In this thesis, I apply genomic epidemiology tools to study the transmission and evolution of emerging viruses with limited sampling. The first study evaluates the evolution of Oropouche virus (OROV) since its emergence in the mid 20th century in South America. The second study describes the transmission dynamics of the first SARS-CoV-2 wave in Ecuador, where genomic sampling was limited. The third study investigates the emergence of a recombinant SARS-CoV-2 lineage in North America and describes its remarkable spread in early 2021. The fourth study compares the spatio-temporal spread of two arboviruses in Mexico, Dengue virus (DENV) and Chikungunya virus (CHIKV) and reveals similar trends in their importation and spread in the country. Finally, the fifth study derives from the investigation of a CHIKV outbreak in Kassala, Sudan, and explores the possible sources of the outbreak.

Acknowledgements

When I was finishing my Master's degree in the University of Edinburgh in 2016, I was encouraged, by my then dissertation supervisor Prof Andrew Rambaut, to apply for a DPhil with Prof Oliver Pybus. I am immeasurably grateful to Oliver for having taken me under his supervision; his knowledge is astounding, and his deep understanding of virus transmission and evolution has been a constant inspiration to me. The COVID-19 pandemic prompted incredible challenges to us all, yet this burden felt remarkably light to me thanks to his support and kindness. He's challenged and mentored me on every step of the way, making me a better scientist and academic as a result.

I am immensely grateful to Dr Marina Escalera-Zamudio, Prof Nuno Faria and Prof Moritz Kraemer, who acted as unofficial supervisors and mentors to me. They have all entrusted me with valuable data and analyses, provided me with ample opportunities to collaborate, do research and gain experience across multiple fields and regions, and looked after me throughout my DPhil. Wherever my academic and professional career takes me, it would have only been possible by your support and trust. I shall not forget it.

A researcher, whether in training or seasoned, is as only as good as their research team and I have been gifted with a group of brilliant, insightful and supportive people to help and guide me through my training. The Evolve.Zoo supergroup has been pivotal to my DPhil, and I am infinitely grateful to Sarah Hill, Louis du Plessis, Julien Théze, Alex Zarebski, Lucy Matkin, Tetyana Vasylyeva, Sabrina Li, Mahan Ghafari, Darlan da Silva Cândido and Jean-Paul Carrera. I arrived at the group as a lone DPhil student, yet having been able to share the journey with Darlan (who arrived a few months after I did) has been invaluable to me. I am so grateful for your support and friendship, and it's been an honour to learn from such a rising star. I'm certain that you'll continue to be an inspiration to me and many others for years to come!

This thesis is a product of multiple collaborations with exceptional teams across the world who have trusted me with data and who have been pivotal to my DPhil. Thanks to my collaborators and friends at USFQ in Ecuador (Paúl Cárdenas, Sully Márquez, Belén Prado, Juan José Guadalupe, and my undergraduate mentors Verónica Barragán and Gabriel Trueba), our collaborators at UKHSA (formerly PHE; especially Hilary Bower, Steven Pullan, and Christopher Logue who inspired me to study viruses many years ago in Ecuador), the InDRE team in Mexico and everyone at the CADDE Project. I am grateful to USFQ and to SENESCYT for their generous financial support during this time.

Life in Oxford has been good as I've had the opportunity to share it with dear friends. Thank you to Tanzil, Matt, Rebecca, Daria, Jacob and everyone in Alan Bullock Close for being the most wonderful community during the uncertain pandemic times. Special mention to Miriam and Catherine, watching over you two beautiful girls has been the most wonderful reminder of what life is all about (thanks Matt and Rebecca for trusting them to us). I could have not done this without my family's love and support, my mom Moni, Sebas, Ana, Francis, Juanpa and Elsa; I hope that I can repay all that

you’ve given during these years and always. Dad, I know you’re watching down on me, and that this dream was as much mine as it is yours. Thank you.
This DPhil adventure, as with all the adventures to come, was shared with my wife, María Mercedes Cobo. I love you and I pray that I never stop fighting to become the best man I can be, for you and for our family. Thank you for taking this leap of faith with me, I am immeasurably proud of your accomplishments and infinitely grateful for your love and support. Finally and ultimately, all gratitude goes to God, the Highest Good to aim for, the “*Logos* rising out of darkness to fill the whole world” (Jonathan Pageau).

Statement of contributions and related publications

The work presented in this thesis was completed between 2018 and 2021 and includes a series of data sets for various emerging viruses from different locations in the Americas and Africa. During 2020, the urgency prompted by the COVID-19 pandemic drove a shift of the focus of my research as I collaborated with the international epidemiological data collection initiative *The Open COVID-19 Data Working Group*, later formalised as the *Global.health* initiative. I also assisted with the setup of the genomic surveillance capacity in Ecuador through collaborations with the CADDE Centre and Universidad San Francisco de Quito. Some of the outputs from these endeavours are included and detailed here.

Chapter 2. Evolutionary Dynamics of Oropouche Virus in South America

I was responsible for the entirety of this chapter. Novel OROV genome sequences from Ecuador were generated by ELW, STP, CEL and GT. TAB, MEZ and OGP supervised the work from conception.

- **Gutierrez B**, Wise EL, Pullan ST, Logue CH, Bowden TA, Escalera-Zamudio M *et al*. 2020. Evolutionary Dynamics of Oropouche Virus in South America. ***Journal of Virology*** 94(5): e01127-19.

Chapter 3. Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador

I was responsible for the entirety of this chapter. Novel SARS-CoV-2 genome sequences from Ecuador were generated by authors from USFQ, UEES, INSPI and Zurita & Zurita Laboratories. Support for the phylogenetic analyses was provided by DdSC, LDP and JTM. PC supervised the data generation, and OGP supervised the manuscript design and all analyses.

- **Gutierrez B**, Márquez S, Prado-Vivar B, Becerra-Wong M, Guadalupe JJ, Candido DS *et al*. 2021. *Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador*. ***Virus Evolution*** 7(2): 1-12.

Chapter 4. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America

I was responsible for the entirety of this chapter. Early work was performed by SF, RH, BJ, CR, MEZ, HC and LD, from which I further developed phylogenetic and recombination analyses conceptualised by AR and OGP.

- Gutierrez B, Castelán Sánchez H, Candido DS, Jackson B, Fleishon S, Ruis C et al. 2021. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. **Cell Host & Microbe** 30: 1112-1123.

Chapter 5. Spatio-temporal trends of arboviral spread in Mexico and Central America

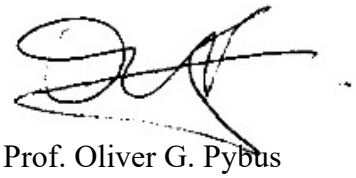
I was responsible for the entirety of this chapter. Novel DENV-2 and CHIKV genome sequences were generated through the GEAPH Workshop in Mexico City in February 2020; the workshop was organised by MEZ with support of NRF and MUGK and dictated by DdSC and BG. Sequencing was performed with collaborators at InDRE: ERMC, APRM, FGA, MdLTR, AAR, CWA, MV and ERG. MUGK, MEZ and OGP supervised the manuscript design and all analyses.

Chapter 6. Phylogenetic analysis of the 2018 epidemic of Chikungunya virus infection in Kassala, Sudan

This chapter presents an investigation of a 2018 CHIKV outbreak in Kassala, Sudan, carried out by the UK-PHRST and the Ministry of Health of Sudan. I contributed to the phylogenetic analyses used to describe the identity and possible sources of the epidemic, and the identification of a potential vector species for the virus.

- Bower H, el Karsany M, Hussein Hussein AAA, Idriss MI, al Zain MA, Alfakiyousif MEA, Mohamed R., Mahmoud I, Albadri O, Mahmoud SAA, Abdalla OI, Eldigail M, Elagib N, Arnold U, **Gutierrez B** *et al*. 2021. *Kankasha in Kassala: a prospective observational cohort study of the clinical characteristics, epidemiology, genetic origin, and chronic impact of the 2018 epidemic of Chikungunya virus infection in Kassala, Sudan*. ***PLoS Neglected Tropical Diseases*** 15(4): e0009387

As Bernardo's supervisor, I confirm that the information presented here is representative of his contributions to the work listed in this thesis.



Prof. Oliver G. Pybus
Professor of Evolution and Infectious Disease
Department of Zoology
University of Oxford
United Kingdom

Contents

1 Introduction

Geographic spread of emerging viruses

Modern conceptions of infectious diseases (such as the One Health approach) recognise the notion that some pathogens circulate between human and animal populations. Under this framework, *emerging pathogens* arises as a category of infectious diseases that have appeared recently in human populations and increased in frequency broadly or within a particular region *(1)*. Emerging viruses are an important category of emerging pathogens and account for a large proportion of such pathogenic microorganisms *(2, 3)*.

Emerging viral pathogens

Viruses are intracellular parasitic entities that, upon infection of host cells, deploy a variety of molecular mechanisms to hijack host cell machinery in order to replicate their genetic material and generate new infectious units called virions or viral particles *(4)*. Viral particles contain genetic material (in the form of DNA or RNA) encased by a

protein capsid *(5)*; they are therefore biological entities and interact with living organisms, although it is debated whether they are classified as living organisms *(6–8)*.

Wildlife, livestock and human viruses

Viruses exhibit considerable variation in their capacity to infect one or more host species. Emerging viruses by definition are capable of infecting different species, but the range of species susceptible to a particular virus is diverse *(3, 9–12)*. The most specialised viruses are constrained to infecting a single species and therefore rely on the infection of new susceptible individuals to sustain their transmission (e.g., *(13, 14)*). More generalist viruses may circulate in at least one 'reservoir' species from which they can 'jump' or spill over into other species. These transmission events between species are called *zoonoses* when occurring from an animal species to humans and can lead to a variety of outcomes, ranging from single human cases with no significant further transmission to other people (e.g., H7N9 Influenza A virus, *(15)*; rabies virus, *(16, 17)*) to small clusters of cases (e.g., Middle Eastern Respiratory Syndrome coronavirus, MERS-CoV, *(18)*), large but transient epidemics (e.g., Severe Acute Respiratory Syndrome, SARS-CoV, *(19)*), and pandemics that can result in the establishment of the virus as a dedicated human pathogen (e.g., human immunodeficiency virus, HIV, *(20, 21)*). Generalist viruses have the capacity to infect multiple species that are often evolutionarily related; for example, rabies virus (*Rhabdoviridae*, RABV) has a wide host range across different carnivore and bat species *(16, 22)*. Similarly, influenza viruses (*Orthomyxoviridae*) circulate freely amongst waterfowl and domesticated birds across at least 26 taxonomic families, predominantly within the orders Anseriformes (ducks,

geese and swans) and Charadriiformes (terns, gulls and shorebirds; *(23)*). Influenza virus strains associated with other vertebrates have been widely studied; these include swine, equine, canine and various types of human lineages, among others *(24, 25)*.

Virions and infection

Viral particles display a high degree of structural diversity, in terms of both particle shape and genome architecture. Broadly, viruses that belong to the same virus order or family share similar virion shape and composition, as well as similar organisation and arrangement of their genomes. Among viruses capable of infecting humans (and other vertebrates), several important viral particle components are shared: *i)* structural proteins, which bind to the viral genome during packaging of new viral particles or form geometrical capsids to protect the genetic material *(5)*, *ii)* an enzyme (either packaged within virions or encoded by the virus genome and produced during cell infection) that assists in the replication of the viral genome during infection, and *iii)* one or more surface proteins that localise on the outermost layer of virions (either attached to an external membrane or directly to the viral capsid) and mediate host cell recognition, attachment and entry during infection *(5)*. Some of these proteins exhibit dynamic molecular evolution due to their exposed nature and can accumulate large numbers of mutations; this occurs as part of an arms-race dynamic that takes place between the host adaptive immune system and the virus *(26)*.

The infection of host cells relies on an initial step where the viral surface proteins directly interact with surface proteins of the host cell, often cell receptor molecules used by the host to regulate cellular and physiological functions *(27, 28)*. Following this initial

interaction, the virus enters or releases its genetic material and necessary accessory proteins into the cytoplasm of the infected cell; this process is mediated by endocytosis or membrane fusion mechanisms *(29)*. The details of the viral replication cycle after gaining cell entry differ widely across viruses but conclude in the assembly and release of new virus particles *(28)*. During this process, the host innate immune system plays a key role in mitigating and blocking viral dissemination. Despite this activity, the co-infection of individual cells with multiple viral particles can sometimes occur and can lead to active genetic exchange between viral particles *(30, 31)*.

Genomic architecture

Viral genomes exhibit substantial diversity in composition and architecture. The major distinction between virus genomes is their chemical composition *(28)*. Virus genomes can be composed of either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and can be configured as single stranded RNA (ssRNA), double stranded RNA (dsRNA), single stranded DNA (ssDNA) and double stranded DNA (dsDNA). The Baltimore classification distinguishes these and considers the mechanisms by which the viral genome generates messenger RNA (mRNA) from its genetic material: this classifies viral genomes into positive sense (where the genome itself functions as mRNA and directly interacts with ribosomes for the translation of genes into proteins) and negative sense genomes (where the genome serves as a template from which mRNA molecules are produced through an RNA-dependent RNA polymerase, RdRp; *(28)*). A separate category is ascribed for viruses with DNA or RNA genomes that utilise reverse transcription.

Another key feature of virus genome structure is the possibility for segmentation, which refers to the number of independent molecules comprising the entire viral genome. Viruses can be classified as segmented and non-segmented depending on whether their genomes are contained within a single molecule or several; in the latter case, each separate molecule is called a chromosome or a segment. The number of segments differs among families; for instance, the *Arenaviridae* family (e.g., Lassa fever mammarenavirus, LASV) includes viruses with genomes comprised of two segments, while *Peribunyaviridae* (e.g., Bunyamwera virus, BUNV; OROV) and *Nairoviridae* (e.g., Crimean-Congo haemorrhagic fever virus, CCHFV) contain three segments and the *Orthomyxoviridae* family (e.g., Influenza viruses) contain six to eight separate segments *(32–36)*. The separation of the genome into several segments facilitates genetic exchange co-infecting viruses through a process called reassortment *(37)*; a related process is also possible for viruses with non-segmented genomes through a process called recombination. In addition to genome segmentation, viruses differ in a series of features that include the presence of overlapping coding regions, untranslated terminal regions (UTRs) and non-coding intergenic regions *(38–41)*.

Transmission dynamics and virus evolution

Virus propagation depends on the generation of new offspring and the colonisation of new environments. As intracellular parasites, this propagation process entails the infection of new, susceptible hosts. The transmission process occurs through the excretion of new infective viral particles, which reach new hosts by different pathways, sometimes through complex infection cycles that include intermediary species. The

most frequent routes of transmission can be classified as either direct transmission between hosts, or vector-mediated transmission. In direct transmission (i.e., where infectious viral particles can infect a new host without the necessity of an intermediate vector), viral particles can be either secreted through the host mucosa (e.g., airways, digestive tract) and reach new hosts through various mechanisms (e.g., aerosols, infectious droplets), or can be transmitted by direct fluid exchange between an infected and a susceptible host. Secreted particles can sometimes survive on surfaces (known as fomites) or other media (such as food and water) and reach the new hosts through indirect contact with the original infected host *(42, 43)*.

Vector-mediated transmission is associated with an intermediary species that becomes infected by the virus and has the capacity to transmit it to other individuals of the original host species. A large number of viruses that infect vertebrate hosts utilise arthropods as vectors (generally classified as *arboviruses*, short for arthropod-borne viruses). Transmission can be either facilitated by the bite of the arthropod vectors during feeding or by contact with the secretions of the vector species *(44)*. The cell and tissue tropism of different viruses is intrinsically associated with their transmission route(s) *(11)*.

The replication of viral genetic material during proliferation inside a host introduces mutations into the genomic sequence. These changes can have effects on gene function and viral fitness and are therefore fundamental to viral evolution. This process is also closely associated with the infection and transmission process because contagion between hosts acts as a genetic bottleneck, such that only a fraction of the genetic

diversity accumulated by the virus infecting an individual host is transmitted to others. These genetic changes accumulate over different timescales (i.e., intra-host variation and between-host variation) and can be tracked to infer the evolutionary trajectories of different viral lineages *(45, 46)*.

Cross-species transmission and zoonosis

The identification of determinants of cross-species transmission is an area of active research most often focused on transmission from non-human to human hosts but increasingly focusing on *zooanthroponosis*, the transmission from human to non-human hosts *(47)*. Both viral and host factors are likely to play important roles in viral zoonosis, and the latter may depend also on environmental factors. The likelihood of zoonoses is associated with the epidemiological dynamics of the pathogen in the non-human population and the frequency at which humans come in contact with this population. Similar factors drive transmission in both non-human and human populations: intrinsic transmissibility of the virus, transmission routes, susceptibility and immunity, population demographics, and individual or social behaviour (sometimes driven by environmental factors, *(48)*). Population size changes can play an important role in zoonoses, as changes in reservoir population sizes and behaviours can drive outbreak risk in human populations (e.g., Lassa fever virus, *Arenaviridae*, LASV, *(49, 50)*). Similarly, the role of vector species also plays a key role; this can include arthropod vectors whose population sizes can vary rapidly as a function of climatic factors *(51, 52)*.

The capacity for cross-species transmission is also heavily influenced by genetic and biological determinants that are intrinsic to the virus. Receptor identification and binding is an important feature of viral host range and tropism *(9, 11)*, and some of these mechanisms have been studied in detail. For example, the capacity of avian influenza A viruses to infect human cells is associated with the hemagglutinin (HA) surface protein's change in specificity from α2-3 sialylated glycans in the avian gastrointestinal tract endothelial tissue to human-like α2-6 sialylated glycans receptors in the upper and lower airways *(53–55)*. More recently, the human angiotensin-converting enzyme 2 (ACE2) protein was identified as the main receptor by which SARS-CoV-2 recognises and infects human cells through its surface Spike protein *(56)*, similar to what had been previously determined for SARS-CoV almost two decades prior *(57)*. Similarities in surface protein structures suggest that SARS-CoV-2 has the potential to infect multiple other mammal species *(58)*, as has been demonstrated by its circulation in mink farms *(59)* and wild white-tailed deer populations *(60)*.

Re-emergence, cryptic circulation and viral reservoirs

While emerging viruses are pathogens that have recently appeared in human populations and/or rapidly expanded their geographic range, re-emerging viruses re-appear in a region after not having been detected for a long period of time, either from re-introductions from other regions or from changes in factors that affect transmission *(61)*. Most emerging viruses of concern to human populations circulate in wildlife populations or in domesticated livestock *(3)* although the frequency at which zoonotic events occur varies between viruses. Under this paradigm, non-human populations

effectively act as reservoirs from which pathogens can re-emerge into human populations when extrinsic factors facilitate zoonotic events *(2, 61–64)*. Viral circulation across scarcely monitored non-human populations can also be conceptualised as a scenario where the pathogen is cryptically transmitted, that is, circulates undetected. This concept is of particular importance for emerging viruses since cryptic circulation often occurs in human populations as well; for example, the emergence of Zika virus (*Flaviviridae*, ZIKV) in the Caribbean and South America was preceded by a period when the virus transmitted undetected in the region *(65)*.

Emerging viruses in the Americas

The Americas span North, Central, South America and the Caribbean Islands, and are an arbovirus-rich region *(66)*. Its climatic and ecosystem diversity and high urban population densities facilitate the transmission of numerous viral pathogens transmitted by arthropod vectors *(67–69)*. While some of the most important emerging arboviruses in the region were introduced to the continent over the last century (such as dengue fever virus, DENV, *Flaviviridae* and Yellow Fever virus, YFV, *Flaviviridae*; *(70, 71)*), multiple viruses endemic to the Americas have the potential to become public health threats, including Oropouche virus (OROV, *Peribunyaviridae*), Mayaro virus (MAYV, *Togaviridae*) and Venezuelan Equine Encephalitis virus (VEEV, *Togaviridae*), among others *(66)*. Furthermore, recently introduced arboviruses like West Nile virus (WNV, *Flaviviridae*, first detected in 1999), Chikungunya virus (CHIKV, *Togaviridae*, first detected in 2013) and ZIKV (first detected in 2014) have also become important pathogens causing noteworthy epidemics in the region *(65, 72–75)*. These viruses are

competently transmitted by mosquitoes of the *Aedes*, *Culex* or *Haemagogus* genus, or to a lesser extent, midges of the *Culicoides* genera. DENV, ZIKV and CHIKV are likely transmitted by *Aedes* mosquitoes associated with urban cycles of transmission, while the YFV transmission cycle in Latin America has been predominantly sylvatic and driven by *Haemogogus* transmission *(71–73, 76)*. Furthermore, clusters of cases and small epidemics of viral haemorrhagic fevers (VHF) have also been reported in recent years in the southern cone of South America, caused by highly pathogenic emerging viruses belonging to the *Arenaviridae* (e.g., Machupo virus, MACV; Junín virus, JUNV; *(77)*) and *Hantaviridae* (e.g., Andes virus, ANDV) families *(78, 79)*. The potential of new viruses to emerge in this region remains high and is linked to the high biodiversity for certain orders (like primates, rodents and bats) in the region, its abundance and the expanding human frontier into previously unaltered ecosystems *(80–83)*. Also, as has been observed with ZIKV and CHIKV, the increasing frequency of travel between locations in the southern hemisphere *(84–88)* rises the risk of new pathogens being exported to other locations.

Respiratory viruses in the Americas follow transmission patterns that are more closely linked to global transportation networks, as has been observed for influenza viruses, for example *(14, 89, 90)*. Transmission dynamics within Latin America are less clear; using influenza viruses as a model, it is likely that transmission lineages there are seeded from North America and Europe *(89–92)* but the region presents its own complex transmission patterns as shown by high rates of vaccine strain mismatches for influenza B viruses across several countries *(93)*.

The emerging viruses studied in this thesis comprise three arboviruses (OROV, CHIKV and DENV) and a respiratory virus, Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2, *Coronaviridae*). Of the first three viruses, CHIKV and DENV are transmitted by *Ae. aegypti* (CHIKV is also transmitted by *Ae. albopictus*), while OROV is believed to be transmitted by *Culicoides paraensis*. They possess either negative-sense (OROV) or positive-sense RNA (DENV, CHIKV, SARS-CoV-2) genomes, as is common amongst emerging viruses.

General concepts of virus epidemiology and evolution

Evolution by natural selection is the process by which organisms change over time as a response to their changing environment. Evolution can be measured in the absence of genetic data, but in the context of this thesis I will focus on molecular evolution. A mutation is a change in the sequence of a genetic code that occurs by biological, physical or chemical processes, and can be inherited by descendants of the individual in which the mutation took place. Mutations generate changes in phenotypes and are therefore susceptible to the effects of natural selection. In viruses, this means that selection acts on the proteins or other molecules that virus genes code for, as well as on other genetic elements that mediate genomic replication or gene expression (for example *(94)*). Mutations occur through a predominantly stochastic process, but the accumulation of changes over generations can result in adaptive evolution if: *i.* the changes are transmitted from progenitors to descendants (heritability) and *ii.* these changes improve the capacity of the individuals to further transmit their genetic information to future

generations. Thus adaptive molecular evolution is contingent on the heritability of mutations and on their functional effects *(95)*.

Viruses are useful models in which to study evolutionary processes, due to their relatively rapid generation times, large population sizes and wide range of evolutionary rates *(96)*. The majority of human emerging viruses accumulate genetic variation at high rates, in part because many have RNA genomes *(12, 97)*. This leads to the evolutionary process taking place in synchronicity with viral spatio-temporal transmission dynamics. Here I delineate general concepts of virus spatial epidemiology and some of the common features that shape evolutionary processes in viruses.

Spatial epidemiology of infectious diseases

Viral epidemics are spatially dynamic phenomena. The consequences of these dynamics are manifold and can account for both diversification (such as the occurrence of genetically structured propagation) and homogenisation (such as the introduction of multiple pathogen lineages into each location).

Quantitative epidemiological methods

Modern infectious disease epidemiology incorporates concepts from population biology and microbiology. While sharing numerous concepts with non-infectious disease epidemiology (such as the role of socio-economic, behavioural and environmental factors in disease occurrence), the nature of pathogen transmission results in a set of measures and models that are unique to infectious disease biology. These

facilitate the quantification and inference of transmission dynamics within a population or sub-populations. Infectious disease epidemiology comprises a vast landscape of methods that lie outside of the scope of this thesis, but some selected key concepts will be discussed below.

In an epidemic the number of infections from an infected individual (primary case) to other susceptible individuals (secondary cases) is summarised as the metric R (sometimes called the reproductive number). The reproductive number can be estimated as an average and is a property of multiple extrinsic factors (environment, connectivity and contact rates between infected individuals, behaviour) and intrinsic factors (inherent transmissibility of the virus, e.g., its natural capacity to be transmitted from an infected to a susceptible host). When measured as a function of time, the reproductive number becomes a function with R_t being the reproductive number R at a time t. This is in contrast to the R_0 parameter, which measures the number of secondary infections produced by a primary case in a population where all individuals are susceptible to contagion (i.e., at the beginning of an epidemic; *(98)*). Values of $R_t > 1$ lead to an increase in epidemic size, while values of $R_t < 1$ lead to a decrease *(99)*. The rate at which the numbers of cases increase or decrease over time is measured as the growth rate and can also be measured as the doubling time.

The fact that hosts can in some cases gain protection against re-infection for a period of time after infection and recovery affects transmission dynamics because susceptible hosts become less common. In reality, this phenomenon can occur over short timespans as other variables change, such as the exhaustion of susceptible individuals in

heterogeneous contact networks or other extrinsic effects that reduce sources of infection (such as infected mosquitoes due to temperature shifts; *(100)*). Epidemiological models represent change in the different categories of individuals: susceptible, infected, infectious, and removed (either through recovery or death). Furthermore, estimation of model parameters can facilitate the forecasting of the trajectory of an epidemic, informing the application of suitable public health measures *(101)*. Accounting for the evolution of the viral pathogen is another key aspect of modelling epidemics caused by fast-evolving viruses *(102)*.

Mapping infectious diseases

Geography can be a strong predictor of infectious disease incidence and contains valuable information about disease spread. Temperature, climatic factors, the distribution of potential sources of disease or vector species, and spatial aspects of human populations are all drivers of infectious disease transmission that are spatially defined. Spatial analysis methods can describe and quantify parameters and provide a strong statistical framework to test hypotheses about infectious disease spread. The geographic resolution at which such analyses are performed is diverse; highly granular data can provide information at a finer scale that can be also aggregated across larger regions. Layered on these site-specific factors, dynamic processes can also be added to spatial frameworks, including human mobility and other dispersion drivers such as migration *(103, 104)*.

Human mobility and infectious disease spread

One of the most important factors influencing infectious disease is the expansion of human populations and the ever-growing connectivity between them. In the context of spatial epidemiology, numerous metrics exist that can account for the flow between geographic locations. Incorporating data on the flow of humans between locations can serve as an important predictor of viral spread *(105)*. Traditional proxy measurements *(106–108)* have been replaced by time-dependent measures that facilitate real-time tracking of human mobility and different levels of resolution. Data on the global aerial transportation network *(109)*, and more recently high-resolution geolocation facilitated by the use of smartphones and other electronic devices can now provide an unprecedented amount of data to finely analyse human movements at highly local, national, regional and global scales *(104)*.

Viral population genetics

The evolution of biological organisms is intrinsically linked to the composition, structure and changes of their populations over time. For viruses, the notion of a 'population' is associated with both a population of infected individuals, and with the genetic diversity of viruses within these hosts. When one infected individual transmits a virus to another only a small portion of the total viral particles circulating in the donor host are transmitted, resulting in a genetic bottleneck *(110)*. Studies of intra-host evolution and transmission bottlenecks for chronic infections such as HIV *(111–113)* and HCV *(114, 115)* have shed light on these processes, but both phenomena have also been described in viruses causing acute infections such as in Dengue virus (*Flaviviridae*,

DENV; *(116)*), Ebola virus (*Filoviridae*, EBOV; *(117)*) LASV *(118)* and influenza viruses *(119–121)*. These inter-host bottlenecks are themselves likely to be subject to evolution *(122)*, and different evolutionary forces could affect the association between virus intra-host evolution and population-level evolution *(123, 124)*.

On a population scale, viruses like influenza virus and DENV exhibit long-term evolution characterised by frequent lineage replacement events; these processes establish a dynamic by which lineages emerge and spread but are subsequently outcompeted, displaced, and potentially even driven to extinction by new emergent lineages *(14, 125–127)*. A similar process has characterised the evolution of SARS-CoV-2, as dominant lineages at one point in time have been displaced by new and emergent variants *(128)*. For tracking purposes, the World Health Organisation (WHO) established a nomenclature for lineages that share a set of noteworthy mutations (also described as a "constellation of mutations"; *(129)*) due to their potential functional effects; these are called variants and classified as Variants of Interest (VOIs) and Variants of Concern (VOCs) for public health purposes. Some of the identified and monitored VOCs have sequentially outcompeted and displaced other lineages at local and global scales, most notably Delta and, as of the time of writing, Omicron *(130, 131)*.

The forces that drive the diversification of viral populations and lineage establishment and displacement at different temporal and geographical scales are manifold and will be discussed in greater detail in Section 1.3. For influenza viruses and SARS-CoV-2, as well as for many other viruses, natural selection results in lineages and variants that display some degree of immune evasive capacity. In these cases, new

dominant viral populations emerge and propagate as they are more capable of efficiently spreading in a population that has acquired immunity to previous lineages or strains *(132)*. Thus the dynamic adaptive immunity of the host population (whether by natural infection or by immunisation) is tightly interlinked with viral population dynamics: increasing immunity in the host population limits viral population sizes, but at the same time exerts a selective force that can drive the propagation of more immune-evasive lineages *(132–134)*. The adaptive evolution of viral populations in face of the ever-changing host adaptive immunity follows a pattern described by the Red Queen Hypothesis (RQH), under which co-evolving populations are under constant pressure to adapt and change to remain viable *(135, 136)*.

Population structure and stratification

The spatio-temporal distribution of viral and host populations plays a major role in shaping the evolution of viruses. The build-up of viral genetic diversity is, for the most part, determined by mutations that are inherited by descendant lineages (vertical gene transfer, from ancestors to descendants) rather than through genetic exchange between individuals, as would be the case for sexually reproducing organisms or prokaryotes engaging in lateral or horizontal gene transfer. This results in virus genetic population structure.

In the context of emerging zoonotic viruses, population structure arises between different host species. The strength of this structuring (i.e., how strictly evolutionary related viruses are confined to a single host species) will be determined by the rate at which cross-species transmission occurs. For example, RABV displays a broad host

range and can easily and frequently jump between host species *(137)*. While some degree of phylogenetic structure is observed for RABV (i.e., lineages infect a single species), both genetic and observational evidence show that RABV crosses the species barrier often in wild and domesticated animal species *(22, 137, 138)*. Other viruses have sustained circulation in non-human hosts and can establish only limited human transmission after jumping into human populations; such is the case for yellow fever virus (*Flaviviridae*, YFV) which circulates widely in non-human primates and causes disease in humans through multiple independent cross-species transmission events *(69, 139–141)* or can establish human transmission that drives large epidemics mediated by one or more intermediate vector species *(69, 142)*.

Another recent example in the *Coronaviridae* family is Middle Eastern respiratory syndrome (MERS-CoV), which circulates widely in camel populations and causes limited but virulent clusters human of disease from independent spillover events *(18)*. Other viruses circulate in a limited range of host species but occasionally engage in cross-species transmission events of high consequence, e.g. some lineages of human, avian, equine and swine influenza A viruses (IAV; *(143)*). Notable events of this kind include various pandemic IAV strains such as the 2009 pandemic H1N1 (H1N1pdm09) strain that originated in pigs *(144–146)*. Species-dependent population structure is observed for pathogens that have circulated in human populations persistently over long periods. A range of pathogens have become highly adapted to, and established in, humans. For instance, DENV and HIV have established persistent circulation in human

populations and are genetically distinct from their closest relatives infecting non-human hosts.

A second dimension of population structure is geographic. As viral pathogens spread and diversify, specific lineages circulate in defined spatial areas, sometimes in isolation. Geographically confined viral lineages accumulate genetic diversity and diversify, leading to the formation of geographically-structured populations or lineages *(147)* that can be identified at the genetic level. Such is the case for many arboviruses, whose lineage designations represent the locations or regions where they circulate or circulated, including CHIKV *(148, 149)*, DENV *(76)*, and ZIKV *(150)*.

Regional variation in climate, ecosystems, biodiversity and the ecology of host and vector species can affect natural selection as viruses adapt to optimise their transmission capacity to the local context. An example of this is the evolution of CHIKV, whose lineages circulating in specific regions have independently acquired a mutation that allows the virus to efficiently infect *Aedes albopictus* mosquitoes as well as *Aedes aegypti*; this mechanism allows CHIKV to spread more efficiently in region were *Ae. albopictus* is more prevalent *(151, 152)*. A similar process drives the evolution of highly pathogenic avian influenza A viruses (HPAI), for which the independent emergence of these HPAI strains can be traced to spatially constrained populations of farmed waterfowl *(153)*. Furthermore, geographic structure can result in spatial variation in host immunity. During the SARS-CoV-2 pandemic, some VOCs saw higher incidences in geographic regions where other variants and lineages were absent or were introduced later in time. Such is the case for VOIs/VOCs such as Beta in South Africa *(154)*,

Gamma in Brazil *(155)*, Lambda in Peru *(156, 157)* and Mu in Colombia and Ecuador *(158, 159)*. In the long term, complex interactions between immune cross-protection and antigenic variation in the virus population can lead to the sustainable establishment of multiple, co-circulating strains *(160)*.

Effective population sizes

Changes in population sizes are closely associated with virus evolution. The large population sizes in viruses, which are a product of the substantial number of infections they can cause and the vast numbers of viral particles within infected hosts, provide very high numbers of opportunities for mutations to occur. In combination with short generation times, virus populations can explore vast areas of mutational space. In population genetics, population sizes are conceptualised as effective population sizes (N_e). This refers to the size of an idealised population that behaves in the same way as a complex, real population in the way in which its genetic composition changes over time *(161)*.

Genome evolution

Evolutionary processes at the population level leave measurable signatures on genetic sequences. These signatures are informative of the functional effects of natural selection on gene and organism function and can serve as tractable markers to explore evolutionary patterns and trajectories across time.

Mutations, substitutions, and evolutionary rates

Mutations occur as a product of both intrinsic features of the genome replication mechanics and the effects of external elements. The enzymes responsible for genome replication introduce random errors in the template copying process; some of these enzymes also have a proof-reading function that allows them to 'identify' and correct errors, but this capacity and efficiency carries across species. For RNA viruses, RNA-dependent polymerases usually don't have a proof-reading function, which makes them more error-prone than DNA viruses.

After a mutation occurs in a genome, it can be passed on to descendants over multiple generations and this trajectory can be tracked. Such mutations are known as substitutions. The number of substitutions that a lineage accumulates over time is measured using a parameter called the substitution rate or evolutionary rate, which depends on the mutation rate, the virus generation time, population size and the effects of selective pressures over time *(162, 163)*.

The types of mutations discussed so far refer to point mutations, which encompass the replacement of one nucleotide for another. However, a second type of mutation is common where one or more nucleotides are either introduced into a genetic sequence or removed from it; these are called insertions and deletions respectively (or jointly, 'indels') *(164)*. Determining the drivers for the distribution of indels is a challenging task *(164, 165)*, but recent analyses of large amounts of genomic data from the well-documented COVID-19 pandemic have shown that similar deletions can occur independently or in parallel on different lineages *(166, 167)*. Molecular mechanisms that

lead to the convergent introduction of insertions on predictable sites have been described in specific viruses; for example, highly pathogenic avian influenza viruses (HPAIVs) can independently acquire a polybasic cleavage site (pCS) in their hemagglutinin (HA) protein due to an RNA editing-like mechanism in the IAV genome *(168)*. The sequences of these pCS are variable but share a sufficiently conserved pattern that renders their functional consequences conserved across lineages *(153)*.

Distribution of substitutions in coding regions

While the genomic distribution of mutations is predominantly random, substitutions are not. This becomes most evident for coding regions; due to the redundancy in the genetic code, non-synonymous or replacement mutations will result in a change in the resulting polypeptide sequence, while synonymous or silent changes will not. Positive and negative selection will, in general, act more strongly on non-synonymous mutations, and thereby result in a non-random distribution of substitutions within virus genomes (often with higher synonymous substitution rates when positive or diversifying selection drives evolution and higher non-synonymous substitution rates when negative or purifying selection drives the process).

This general pattern can be affected by other factors. Viral genomes often feature overlapping open reading frames (ORF, genetic sequences presumed to encode for a gene product given that they begin with a ‘start’ signal codon and end with an ‘end’ signal codon), given that considerable amounts of genetic information need to be encoded in their generally small genomes *(39)*. Viral RNA genomes also partially fold into secondary RNA structures *(169–171)*; this also affects the distribution of

substitutions in the genome as disturbing functionally relevant structures would affect viral fitness *(40, 94, 172)*.

Recombination and reassortment

Virus recombination is a process by which genetic material from two genetically distinct parental viruses is combined into a viable descendant. For virus recombination to occur, two viruses need to co-infect the same host cell, a scenario that provides the setting during which chimeric genotypes can emerge. The composition of the chimeric genome usually occurs through molecular processes such as template switching (where the polymerase disengages from the template molecule during replication and re-engages with a different template) or homologous recombination (whereby genetic material is exchanged between two molecules mediated by the host cell's molecular machinery; *(173)*). The resulting molecule is packaged into new virions and, if viable, can be released from the host cell and infect new individuals; the resulting progeny occasionally spreads through the host population. A similar process called reassortment occurs with viruses with segmented genomes; in these cases, genome segments from different co-infecting parental viruses are replicated as individual molecules which are then packaged into virions where different segments correspond to a combination of the parental viruses *(173)*.

These processes can have significant evolutionary consequences. Genetic exchange of this kind has been shown to facilitate the adaptation to changing environments in viruses under experimental conditions (e.g. poliovirus; *(174)*), and has been observed to occur within hosts during prolonged infections (e.g. HIV; *(175)*). At the population

level, it is believed that reassortment in seasonal IAVs in humans assists in the adaptation to the host population adaptive immunity *(176, 177)*; also, a recombination event is associated with the emergence of a MERS-CoV lineage that became dominant in camels in the Middle East between 2014 and 2015 *(178)*.

Phylogenetics, phylodynamics and genomic epidemiology of viral pathogens

Genetics and evolutionary theory have a long history of being used to study infectious disease outbreaks *(132, 179–181)*. Molecular markers (before genetic sequencing became commonplace) were used to establish links between pathogens causing outbreaks and helped to describe and reconstruct epidemics *(182–184)*. With the advent of genomic sequencing, the ability to establish these epidemiological links increased and its resolution improved. The use of phylogenetic methods that were traditionally used to explore the evolutionary relationships between species became an important tool for understanding pathogens in an evolutionary and epidemiological context.

Given that some viruses (particularly RNA viruses) evolve at rates several orders of magnitude faster than some prokaryotes and most eukaryotes *(96)*, the substitutions that accumulate during their transmission during an outbreak can be observed at the same timescale as they are being sampled. Virus genomes also contain information regarding spatial movements, the temporal distribution of transmission events, and genetic changes. Data sets in which genetic variation has accumulated over the same timescale

as the process of sequence sampling are called heterochronous data sets *(185)* and the involved populations are referred to as measurably evolving populations (MEPs).

Phylogenetic methods

Phylogenetics is the field of evolutionary biology that deals with the reconstruction of the evolutionary history of a set of taxa. While traditionally taxa refer to different species related by descent from common ancestors, the development of high throughput sequencing means that taxa can represent individuals rather than species, facilitating the exploration of intra-species evolution. Modern phylogenetics uses genetic and genomic data as the information source for evolutionary inferences, and its methods have evolved with time as genetic and genomic data sources became more readily available and computational power has grown to meet the greater demands of large data sets *(186)*.

Phylogenetic trees represent sequence data from individual taxa as tree tips (also called tree leaves) which are connected through lines conveying descent from a common ancestor, called branches. The length of branches can represent a measure of genetic change (e.g., number of expected substitutions per site) or time, if the phylogeny is calibrated through the use of molecular clock models. Internal nodes represent common ancestors, and all branches can be traced back to a node that is ancestral to all taxa, called the tree root (not all trees are inherently rooted as most methods need an external information source to inform the root position). For n taxa, there are $(2n - 3)!/2^{n-2}(n-2)!$ potential phylogenetic trees that represent their putative evolutionary trajectories. The purpose of molecular phylogenetics is to infer the tree or trees that best explain the observed set of taxa by comparing and ranking competing hypotheses (each potential

phylogenetic tree acts as a hypothesis). Selecting the most adequate tree given the data can lead to multiple phylogenetic tree shapes (also called topologies) being equally likely explain the relationships between genetic sequences *(187)*; therefore, an important element of phylogenetic inference is the representation of uncertainty across phylogenies, whether through the tree inference process (such as with Bayesian phylogenetics) or by approximating a null distribution against which to compare the inferred tree (e.g., bootstrapping *(188)*, approximate likelihood ratio tests; aLRTs; *(189)*).

The simplest molecular phylogenetic methods, which I won't discuss in detail, use algorithms to quantify sequence variation among taxa and propose hierarchical clustering of the sequences, or sought to identify the phylogeny that was most parsimonious in explaining the evolutionary relationships between the included samples *(190)*. The phylogenetic methods mostly employed in this thesis are based on either Maximum Likelihood (ML) or Bayesian approaches.

Maximum likelihood phylogenetics – heuristics and inference

Maximum Likelihood (ML) phylogenetics employs the probabilistic concept of *likelihood* as the optimality criterion, and to identify the tree that best explains the data. ML methods attempt to identify the tree with the maximum likelihood (MLE), conditional on the genetic data, amongst all the numerous possible trees. The ML framework also explicitly assumes an underlying sequence substitution model that governs the process under which the genetic data was produced. These probabilistic nucleotide substitution models are formulated as Markov processes that define the

equilibrium frequency at which individual bases occur (π) and the relative rates at which they change (μ). Models have been developed that incorporate different numbers of parameters to account for different rates and frequencies, from the simple Jukes-Cantor (JC) model (a single parameter for π and a single parameter for μ) to the highly parametrised General Time Reversible (GTR) model (individual rates for every base frequency and every bi-directional rate of change from each base type to all others; *(191)*).

Heuristic approaches are used to explore and evaluate tree space (the set of potential tree topologies that could explain the genetic data). In ML phylogenetics, heuristic search begins with a starting tree (which can be a random tree or a simple distance-based tree). Next operations are performed to alter its topology, called branch swapping operations. The resulting altered tree is evaluated to calculate its likelihood conditional on the substitution model and genetic data. This process is then repeated in an iterative manner to optimise the likelihood until the MLE is achieved.

ML has become one of the most widely used molecular phylogenetic inference methods due to its moderate computational requirements and relatively few user choices when performing an analysis. The selection of substitution model is one of the main considerations with this method, as different models are better suited for different data sets. Given that nucleotide substitution models are a series of nested models with different number of π and μ parameters (e.g., JC is a special case of the GTR model where all π and μ parameters adopt the same values), model-testing through hierarchical likelihood ratio tests (LRT) is an adequate approach to identify models that best fit the

data; this is the case as the ratio between the maximum likelihood estimate for the null hypothesis (more simple, less parametrised model) and the alternative hypothesis (more complex model with more parameters), called the Λ statistic, is distributed as χ^2 with q degrees of freedom *(190, 192)*.

Bayesian phylogenetics – heuristics and inference

Bayesian phylogenetics implements Bayes rule to estimate the probability that an individual tree from the tree space is plausibly true given the data,

$$Pr(\tau|G) = \frac{Pr(G|\tau)xPr(\tau)}{Pr(G)}$$

where the probability of a tree τ_i given the genetic data G, $Pr(\tau_i|G)$ (called the *posterior probability*), is a function of the conditional probability of observing that data given that the tree is true ($Pr(G|\tau_i)$, also called the *likelihood function*), the probability that the tree is correct before the data G was considered ($Pr(\tau_i)$, called the *prior*) and the probability of observing the data G marginalised by all possible tree topologies ($Pr(G)$, called the *marginal probability*). In this formulation, a probability is assigned to the data (given a hypothesis) and not only to the model parameters *(193)*. This has two important implications for practical Bayesian phylogenetics. On one hand, the users are required to incorporate some prior beliefs about the tree as a *prior*; these are incorporated into the Bayesian framework in the form of probability density functions (PDF) of specific model parameters and the tree shape (defined as a tree prior). On the other hand, phylogenetic uncertainty is inherently estimated from the statistical inference process given that no single tree is taken as the solution but rather a *posterior distribution* of

plausible trees. This sample of trees may contain multiple topologies that represent the probability that certain tips cluster together. Another important property of the Bayesian formulation is that the posterior is conditioned on the genetic data and on various model parameters; this allows for the basic framework to be expanded by incorporating additional underlying models beyond the nucleotide substitution model, such as demographic models and molecular clock models (which can also be applied to ML approaches).

Under most implementations, Bayesian phylogenetics sample tree space through the use of the stochastic Metropolis-Hastings algorithm, which is a variation of the Markov chain Monte Carlo (MCMC) method. This algorithm explores tree space by estimating the *posterior probability* of two different trees sampled from tree space and dividing these values to obtain an acceptance ratio. MCMC is an iterative process: depending on the value of the ratio, the algorithm is likely to accept or reject each 'step' and repeat the procedure of sampling a new tree from tree space and estimating the acceptance ratio. Ultimately, the sequence of steps the MCMC takes will tend to visit regions of tree space with higher posterior probabilities more frequently. In this way, after many iterations, the algorithm records a sample of trees that is an approximation of the posterior tree distribution. This distribution is normally summarised as a single phylogeny such as a Maximum Clade Credibility (MCC) tree: a representative phylogenetic tree from the posterior tree sample which corresponds to the tree with the maximum *a posteriori* (MAP) topology. An MCC tree is estimable because the posterior probability of finding specific clades among the tree sample is additive (i.e., the MCC is the tree where the

sum of the posterior probabilities of finding the clades contained within it is maximised among all other trees from the sample); these posterior probability values are assigned to nodes in the tree and act as an estimate of the uncertainty for individual clades *(194)*.

Viral phylodynamics

The application of phylogenetic methods to genetic data sets from fast-evolving viruses allows for the joint exploration of different evolutionary, spatial and demographic processes, and how they are affected by epidemiological, environmental and host immunity factors. These make up the field of phylodynamics, which has become an important tool in genomic epidemiology *(180)*.

Time-calibrated phylogenetics of rapid-evolving viruses

The calibration of a phylogenetic tree so that branches are proportional to time elapsed between nodes is possible due to the implementation of molecular clock models *(195)* which establish an association between the accumulation of genetic change and time *(196)*. The accumulation of substitutions across time at a constant rate results in a strict molecular clock model *(197)*; however, this simplified model doesn't account for some of the complexities of sequence evolution, both at the mechanistic and population levels (e.g., positive and negative selection by environmental factors that affect specific lineages; population size expansions and contractions, the limiting 'mutation carrying capacity' of genomes that leads to saturation and the loss of temporal signal from the data; *(162, 198)*). To address these phenomena, different molecular clock models have been developed that provide additional parameters to account for variation of

substitution rates and give the users options to apply a model that best fits a particular data set. A more flexible model is the relaxed molecular clock, where a statistical distribution (commonly defined as a lognormal distribution) is used to account for the variation in rates across branches in a phylogeny; *(199)*). Other clock model can capture phenomena such as rate time-dependency *(200)* and phylogenetic dependency *(201)*.

The identification of heterochronous data sets begins by establishing the existence of an association between genetic divergence and elapsed sampling time. This is commonly achieved by plotting a regression between the sum of the lengths of the branches that connect a tree tip with the tree root (also known as the distance from the root) and the sampling date of said tip across all taxa in a phylogeny. A strong correlation between both parameters is interpreted as a robust 'temporal signal' in the data set (also referred to as 'clock-like behaviour'; *(202)*). This approach does not provide a null distribution to perform formal hypothesis testing; alternatively, other statistical frameworks perform permutation tests by randomising sample collection dates across the tree tips and estimating the substitution rates from the data sets to approximate a null distribution against which the real data set can be compared against *(203)*.

Time-calibrated phylogenetic trees for heterochronous data sets (with sampling times for all or most taxa) can be inferred using both ML *(204)* and Bayesian approaches *(205–209)*. Frequently used ML approaches such as the TreeTime implementation *(210)* usually operate on fixed tree topologies and therefore function as a branch length-rescaling procedure, where the node ages are a function of the branch length probability distribution and conditional on the sequences assigned to the node and its descendant

branches. On the other hand, Bayesian methods such as those implemented in BEAST *(207–209)* allow for the co-estimation of the molecular clock parameters jointly with other model parameters and the tree topology. Similar to the Bayesian estimation of phylogenetic uncertainty, Bayesian methods provide an empirical posterior distribution of the molecular clock parameters and node neights which can be used to ascertain the uncertainty associated with node ages *(195, 199)*.

Inferring spatial spread through phylogenetic methods

Phylogenetics have long been associated with the spatial distribution of the species being investigated, as geographic population spread can occur simultaneously with genomic evolution. This has led to the emergence of the field of phylogeography, which incorporates spatial information of sample collection locations to infer the geographical spread of taxa *(211)*. In viral phylodynamics, phylogeographic methods can be useful tools to evaluate hypotheses of how viruses spread in two-dimensional space and to explore the potential drivers for these patterns *(14, 179, 212)*.

Modern phylogeographic methods applied to viral data can be classified into two groups: spatial diffusion approaches and population genetics approaches (not discussed here). The first approach models geographic dispersal as a discrete or continuous diffusion process along the phylogeny which is dependent on the inferred location of ancestors (i.e., nodes) within the tree. Ancestral locations can be inferred under the spatial diffusion model using continuous-time Markov chains (CTMC) to infer the changes between states through time. On the other hand, the second approach is focused

on modelling population-level dynamics and their interplay with spatial mobility framed as a migration rate between the populations *(211, 213–215)*.

Among the spatial diffusion approaches, discrete phylogeographic methods treat sampling locations as discrete traits assigned to individual tree tips and proceeds by inferring the transitions between discrete states across the phylogeny *(216, 217)*. These methods operate as a discrete trait analysis (DTA) in which inferred node locations define the realisation of transitions between states. Therefore, the parametrisation of these models depends on the number of different sampling locations that are coded as distinct discrete states. Depending on sampling granularity and the level of resolution at which different locations are discretised (e.g., samples can be discretised as being collected in the same municipality, city, province, country or region), DTA methods will be defined by the number of states (i.e. different sampling locations) in the model specification; the resulting number of parameters (incorporated in an instantaneous transition matrix for the phylogeny) depends on whether the model is formulated as symmetric (e.g., $S_iS_j = S_jS_i$) or asymmetric (e.g., $S_iS_j \neq S_jS_i$). One of the consequences of this approach is that node locations can only be inferred to have occurred in locations where tree tips were sampled from *(211)*. As an alternative to the DTA methods, continuous phylogeographic diffusion methods utilise the individual geographical coordinates at which individual sequences are collected and model the spatial diffusion process as a form of Brownian process. As such, tree nodes can be inferred to have occurred at unsampled sets of coordinates resulting from random walk models operating within the phylogenetic structure *(218, 219)*.

Bayesian phylogenetic frameworks facilitate the incorporation of non-genomic data and can therefore provide a platform to evaluate the effects of various co-variates on the spatial diffusion process *(211)*. The most widely used implementation of this form of hypothesis-testing parametrises each rate of transition between locations as a function of various predictors and formulates a generalised linear model (GLM) to evaluate the contribution of each potential predictor on geographic spread *(14, 212)*. Furthermore, the discrimination of specific transition rates between locations (usually from a DTA analysis) that significantly explain the general diffusion process can be identified using a Bayesian stochastic search variable selection (BSSVS) approach *(217)* which provides a formal Bayes factor (BF) test to estimate the significance of the contribution of each transition rate in the model, and its posterior probability (PP) of inclusion. The temporal distribution of the movements between locations, while not being directly observed, can also be reconstructed given the Markov properties of the CTMC. This is accomplished by tracking the expected transitions between states (and the opposing waiting times) along tree branches using a robust counting method *(220)*.

Coalescent models and demographic changes across time

An association between the genetic diversity of a population and its effective population size can be established through a theoretical framework known as coalescent theory *(221)*. The coalescent model establishes a relation between the time it would take for two sequences from a population to be tracked back to a common ancestor, and the size of the population. Coalescent models have been expanded to accommodate scenarios where population sizes change across time (e.g., *(222, 223)*). The family of

coalescent models provides an expectation for the accumulation of genetic diversity given a particular demographic history, and therefore makes it possible to explore the role of other forces on the evolution of a population.

Changes in the size of a population as it evolves can have profound effects on the shape of phylogenetic trees inferred from genetic sequences *(224, 225)*. The association between common ancestry and population size established through coalescent theory establishes an inverse correlation between effective population sizes (N_e) and the probability of two lineages in a phylogeny converging after a given number of generations *(221)*. This model, originally defined for populations of constant size, was later expanded to account for changes in population size across time *(222, 226)* and fostered the development of inferential methods for demographic changes back in time from heterochronous data sets *(227)*. The effect of changing population sizes on the general shape of a phylogeny make the joint inference of population histories and substitution rates an important aspect of modern phylodynamics *(228)*.

Incorporating coalescent-based methods to infer changes in demographic histories requires the inclusion of demographic models, which are mathematical functions that describe the change in population size across time *(227)*. Amongst these models, the simpler ones parametrise population growth or decay as a function in which one or a few parameters define the demographic dynamics across the entire phylogeny; this includes constant population size or population growth models (where changes in effective population times are defined by an exponential or logistic function; *(229)*). While simple models are sufficient for some data sets, more complex demographic

trajectories may not be accurately represented by a simple function. For these scenarios, a series of non-parametric piecewise models have been developed which can account for different effective population size estimates at different points in time *(230)*.

For heterochronous data sets, the skyline family of models take different approaches to defining the points in phylogenetic time where N_e estimates are produced, while incorporating phylogenetic uncertainty under a Bayesian framework *(230)*. The classic skyline models *(223, 231)* and the Bayesian skyline plot *(227)* estimate N_e across inter-coalescent intervals defined by a fixed tree topology for the former, and across a set of trees for the latter. The skyride *(232)* expands the basic skyline method and incorporates the intervals between sampling events, and the skygrid model *(233, 234)* allows the user to specify a set number of timepoints at which N_e will be estimated. Depending on the number of taxa, the skygrid can result in a reduction of model complexity and can yield more directly interpretable results depending on the selection of timepoints (e.g., selecting timepoints that represent epidemiological weeks of a sampled outbreak lends to the interpretation of changes in population sizes as proxies for changes in the outbreak epidemiological trends *(230)*. Precisely in this sense, the skyline models can be applied to infer past changes in population sizes independently from other conventional sources of information such as those derived from epidemiological data. However, it has also become common practice to parametrise phylogenetic models to account for these changes in population size as a means to improve the accuracy of other parameters of interest, such as substitution rates. In these instances, the N_e estimates act more as nuisance parameters in the phylogenetic inference.

Evaluating selection on genomic data

The patterns of nucleotide substitution across different regions of a gene or genome, coupled with population genetics theory, make phylogenies a viable hypothesis-testing framework for the effects of natural selection on genetic sequences and phylogenies. The neutral theory of molecular evolution proposes that the majority of polymorphisms observed in a set of sequences have no measurable effect on fitness *(235, 236)*. For coding regions, third codon positions represent a natural site category that is more likely to comply with the neutral theory because mutations on these sites are much more likely to result in a synonymous change *(237–239)*. This expectation can provide an approximation to a neutral distribution for substitution rates against which other site categories can be compared. In a simplified scenario, negative selection eliminates amino acid-altering substitutions (as they are likely to reduce fitness) while positive selection favours amino acid-altering substitutions (as they are likely to be beneficial and would be selected for).

The basic rationale of comparing the substitution rate of sites that lead to non-synonymous changes (d_N) versus synonymous changes (d_S) is used as the basis for phylogenetic hypothesis-testing frameworks that evaluate the likelihood of different modes of natural selection operating on a group of taxa. Broadly, such "selection tests" use estimates of the ratio of synonymous and non-synonymous substitution rates (d_N/d_S). Methods based on the d_N/d_S ratio have become prevalent in the study of virus evolution *(240–242)*. Given that natural selection can operate on specific regions of a gene or genome and on specific lineages within a phylogeny (rather than the entire phylogeny),

these models for selection can be classified as site models (which evaluate the effects of natural selection on individual codons across a phylogeny), branch models (which evaluate the effects of natural selection on individual lineages within a phylogeny) and branch-site models (which evaluate the effects of natural selection on some sites for some lineages across a phylogeny; *(243)*).

Different implementations of these tests have been made available under ML and Bayesian frameworks. As with other aspects of viral phylogenetics, ML implementations operate on fixed tree topologies, while Bayesian methods are implemented within the Bayesian inference framework. Given the dynamic nature of natural selection, different implementations focus on evaluating the signal of different natural selection regimes being either pervasive *(244, 245)* or episodic *(246, 247)* across a data set. Within the Bayesian framework, the inference of site-specific d_N/d_S ratios can be achieved through the implementation of robust counting *(248)*.

*Using parallel evolution as a tool to study virus evolution**

The repeated occurrence of similar or identical genetic changes in different viral lineages is an evolutionary scenario that can be effectively analysed through molecular phylogenetics. Parallel or convergent molecular evolution is the independent emergence of the same genotype or phenotype from distinct ancestors. The relative simplicity of viral genomes and their rapid evolution make them useful models for studying parallel evolution by natural selection, and in some instances represent epidemiologically significant events themselves. Parallel evolution is often adaptive and occurs as a response to similar selective pressures from the environment. Some noteworthy

instances of parallel evolution include the repeated emergence of specific mutations associated to HIV drug resistance *(296)* or antigenic escape *(297)*, multiple zoonotic events associated with the same genetic change in HIV *(298, 299)* and RABV *(300)*, the acquisition of glycosylation sites in H7 subtype AIVs associated to cross-species transmission *(301, 302)* and the repeated acquisition of an E protein mutation that is adaptive to specific vector species in CHIKV *(151)*, among others.

Given that parallel molecular evolution entails observing the same genetic change independently occurring multiple times, it can be conceived as the replication of an evolutionary experiment. By considering the explicit phylogeny that defines the evolutionary trajectories between taxa and inferring the ancestral state where such mutations were most likely to have occurred, a pipeline can be proposed to study these events: lineages where convergent changes occur can be tested for signals of natural selection as a proxy for potential adaptive changes, and the location of these mutations or genetic features can be mapped to resolved atomic structures of proteins. This framework provides a robust analytical framework to explore the effects of convergent mutations which can be confirmed experimentally.

Evaluating reassortment and recombination on viral genomic data

The notion that a single phylogenetic tree explains the evolutionary history of a set of genome sequences relies on the assumption that these sites, while evolving independently, are jointly co-inherited across generations. While this is true in most cases, notable exceptions are viruses with high rates of genomic recombination *(173, 249)* or those with segmented genomes which engage in frequent reassortment *(250,*

251). In these instances, a phylogeny for the entire genome would not represent the true evolutionary trajectory of the complete set of sequences. Therefore, a common practice in viral phylodynamics is to perform tests that evaluate the occurrence of evolutionary discrepancies *a priori* to identify and remove sequences that are potential recombinants, or to explore these recombination patterns and their role in viral evolution.

One of the basic criteria to determine the possibility for recombination is topological incongruency: discrepancies between the shapes of phylogenetic trees for partial genome segments from the same set of taxa *(252, 253)*. This approach is straightforward to implement for viruses with genomes that are naturally segmented, for which the standard practice is to infer a separate phylogeny individually for each segment *(176, 254–256)*. However, there are many other viral pathogens for which recombination is present or pervasive despite their genomes being comprised of a single molecule. In these scenarios, an additional challenge lies in the determination of the point across the length of the viral genome where the recombination event occurred. Inferring these recombination breakpoints can be achieved through different methods; these range from the estimation of pairwise genetic differences of overlapping, fixed-width windows that span the entirety of the viral genome *(252, 257, 258)* to the implementation of a heuristic algorithm to explore possible recombination breakpoints across an alignment, assign relative support to each, and perform a model support comparisons to establish the likelihood of multiple independent phylogenies *(259, 260)*.

Use of phylodynamic tools to investigate viral outbreaks and epidemics

The phylodynamic toolbox

The set of phylodynamic tools described here has been routinely applied to the investigation of viral outbreaks in recent years. The feasibility of their usage stems from the increasing capacity to sequence complete viral genomes directly from clinical samples in near real time and the advances in both the methodological implementations and computational capacity for statistical inference *(179, 261–264)*. The most prominent recent examples of the application of these tools include the H1N1pdm09 influenza pandemic *(145, 265)*, the SARS-CoV epidemic in China and other countries *(266–270)*, the EBOV epidemic in Western Africa *(271–275)*, the ZIKV epidemics in the Americas *(65, 73, 262, 276)*, and more recently the SARS-CoV-2 pandemic *(154, 277–282)*. However, much ongoing research places a strong emphasis on the retrospective analysis of past epidemics, their driving forces, and the long-term evolutionary processes that shape numerous endemic, seasonal, emerging and re-emerging viruses *(84, 141, 283–286)*. On the other hand, the notion of applying phylodynamics as a prospective tool will likely require further theoretical and methodological development given the challenges of accounting for the myriad of factors that drive the evolution of viral pathogens and their simultaneous spread across space.

Similar analyses can be applied to address key questions concerning viral transmission and evolution whether dealing with either real-time outbreak investigation or retrospective analyses *(179)*. Outbreak investigations initially rely on the identification of the causal pathogen, a task which has been efficiently tackled with

molecular techniques and more recently metagenomic sequencing *(287, 288)*. Upon establishing the identity of the viral pathogen and generation of genomic data (often from clinical specimens), the application of molecular clock phylogenetic models and phylogeographic methods can inform on the source of an outbreak (or circulating viral diversity) and describe its geographical spread. This can include both the location where the outbreak started and, in the case of zoonotic pathogens and spillover events, the species where the virus circulated before it sparked the epidemic *(146, 265, 269)*. While genomic data facilitates this type of investigation, the application of these methods becomes particularly insightful when it is used in combination with standard epidemiological data. The spatial transmission patterns are another important aspect of an outbreak which can be investigated using genomic data to describe the circulation of specific lineages across different locations and regions; this and can be combined with demographic and environmental data to evaluate extrinsic factors driving transmission. Examples of this include the determination of the number of independent introductions of a pathogen into a population (which can be an important factor in epidemiological modelling *(289)*) and the identification of key factors that drive transmission following virus importations *(212, 273)*. Regarding viral behaviour and evolution, actively investigating the signatures of genetic evolution can provide insights into the selective forces affecting a virus, suggest functional roles of different genetic elements and inform possible scenarios of the future behaviour of an epidemic *(290)*.

Genomic epidemiology pipelines provide powerful frameworks to explore the transmission dynamics of viruses and provide a path to inferring the occurrence of events

for which no epidemiological data was collected. However, genomic sampling represents a challenge in of itself, as phylodynamic inference can lead to biased estimates when important spatiotemporal gaps in sample collection are present *(214, 291)*. These situations are pervasive among genomic data sets and require careful consideration and potentially further methodological development to improve the inferential power and accuracy of the phylodynamic toolbox (e.g., *(292–295)*).

Thesis scope and outline

Incomplete sampling is likely a pervasive phenomenon among genomic data sets analysed using modern phylodynamics. While this can manifest limitations in their precision and accuracy, it is also noteworthy that the signatures of these sampling gaps can become evident in the results from such analyses, and can be used as a tool to effectively interpret phylogenetic and phylodynamic results and to even gain insights into both the sampling and evolutionary processes that produce a given data set.

In this thesis, I use a genomic epidemiology approach to investigate the transmission dynamics, geographic spread and molecular evolution of various emerging viruses in different settings where genomic sampling coverage is limited. I look at three different arboviruses and one emerging pandemic respiratory virus as they circulate in various regions across the Americas and Africa and apply a diverse set of evolutionary tools to understand how these circulated and evolved over different timescales. I propose that hypotheses can be formulated and tested from specific patterns observed in phylogenetic analyses that reflect the sampling profiles that these viruses display in their respective contexts. Combined with information derived from non-genomic data sources, I

highlight specific insights that can be drawn into the particular sampling gaps across the virus evolutionary history and some of the processes which took place in these sampling blind spots.

In **Chapter 2** I present an analysis of the evolutionary dynamics of OROV, an American Peribunyavirus associated with numerous outbreaks in the Amazon basin, predominantly in Brazil and most recently in Peru and Ecuador *(304, 305)*. My analysis incorporates (at the time) new OROV complete genome sequences and a publicly available data set (spanning the three viral genome segments) to perform a phylogenetic reconstruction of the virus evolutionary history. These analyses reveal a markedly distinct evolutionary trajectory for the viral M genome segments; through a combination of molecular clock analyses and the examination of incongruent tree topologies, I conclude that unsampled viral diversity would account for vastly differing genome segment ages. I also perform a set of selection analyses coupled with structural modelling and reveal that adaptive selection shaped the evolution of the viral Gc spike protein, and that novel lineage-specific N-linked glycosylation sites define two distinct lineages of the M segment.

In **Chapter 3** I perform a phylogenetic analysis of the first year of the COVID-19 pandemic in Ecuador, South America, by analysing all genomes generated by the SARS-CoV-2 genomic surveillance efforts in the country during 2020. This analysis describes and quantifies the number of independent introductions of the virus into the country, describes the resulting transmission lineages and tracks their domestic circulation. While the genomic sampling coverage was unprecedentedly high in the country, the relatively

low numbers of sequences in relation to the estimated cases in the country (derived from excess mortality data) show that the dramatic first wave of the epidemic was scarcely sampled. This is reflected by the time calibrated phylogenies and the discrete phylogeographic reconstructions of the virus. However, evidence for possible ports of entry and the role of the largest cities in the country in seeding the epidemic across different provinces is observable and provide an insight into the transmission patterns of SARS-CoV-2 in Ecuador.

In **Chapter 4**, I present an investigation that led to the identification of a recombinant SARS-CoV-2 lineage that emerged in North America and circulated for a prolonged time in the region, predominantly between Mexico and the USA but also in Central America and Europe. Through the use of recombination analyses and mapping the occurrence of deletions across the evolutionary trajectories of four SARS-CoV-2 lineages (B.1.627, B.1.628, B.1.631 and B.1.634) I identify that the most widespread of these lineages, B.1.628, emerged from a recombination event between B.1.627 and B.1.634. This investigation resulted in the re-assignment of B.1.628 as XB under the Pango nomenclature *(306)*, becoming the second recombinant lineage in this naming system after the XA lineage in the UK *(307)*.

Chapter 5 analyses a historical data set of two arboviruses circulating in Mexico and Central America, CHIKV and DENV. The analyses contrasts some of the temporal and geographical spread patterns of these viruses which, despite showing notorious differences in their seasonal dynamics in the country (CHIKV caused a large epidemic in 2015 while DENV has been established as a seasonal virus in the country for over

three decades), are transmitted by at least one shared vector species, *Ae. aegypti*. By performing phylogeographic analyses and incorporating epidemiological data (collected by the National Epidemiological Surveillance System, SINAVE) and anonymised human mobility data (collected from smartphone usage), I find that the epidemiological dynamics of both viruses are driven by new importations. I also show that the southwest region of the country acts as the main port of entry for arboviruses from various locations, and that viral spread within the country follows a repeatable pattern.

Finally, in **Chapter 6** I present a simple phylogenetic analysis of a CHIKV outbreak in Kassala, Sudan, as part of a larger investigation by the UK Public Health Rapid Support Teams (UK-PHRST) and the Ministry of Health. My contribution to this investigation focused on the phylogenetic analysis of 120 complete genome sequences in context of the broader CHIKV genetic diversity. I find that the outbreak was caused by a single introduction of the ECSA genotype, particularly of viruses from the Indian Ocean Lineage (IOL). This represents a re-introduction of the IOL into Africa after its original departure from the continent into the Indian Ocean territories, India and the Middle East in the early 2000s. Evidence for various unsampled CHIKV outbreaks in the region suggest that this introduction would follow directly from India, where the virus circulates substantially, or from more recent unsampled outbreaks in the Middle East. A reconstruction of the A226V mutation in the viral E protein shows that this variant of the virus is likely transmitted by *Ae. aegypti* rather than *Ae. albopictus* (as was observed in previous outbreaks in the Indian Ocean).

References

1. N. D. Wolfe, C. P. Dunavan, J. Diamond, Origins of major human infectious diseases, *Nature* **447**, 279–283 (2007).
2. H. K. Alexander, T. Day, Risk factors for the evolutionary emergence of pathogens, *J. R. Soc. Interface* **7**, 1455–1474 (2010).
3. L. H. Taylor, S. M. Latham, M. E. J. Woolhouse, Risk factors for human disease emergence, *Philos. Trans. R. Soc. B Biol. Sci.* **356**, 983–989 (2001).
4. E. Domingo, C. R. Parrish, J. J. Holland, *Origin and Evolution of Viruses* (Elsevier, ed. 2, 2008).
5. M. G. Rossmann, Structure of viruses: A short history, *Q. Rev. Biophys.* **46**, 133–180 (2013).
6. P. Forterre, Defining Life: The Virus Viewpoint, *Orig. Life Evol. Biosph.* **40**, 151–160 (2010).
7. E. V. Koonin, P. Starokadomskyy, Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question, *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 125–134 (2016).
8. P. López-García, D. Moreira, Viruses in Biology, *Evol. Educ. Outreach* **5**, 389–398 (2012).
9. M. Woolhouse, F. Scott, Z. Hudson, R. Howey, M. Chase-Topping, Human viruses: Discovery and emeraence, *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2864–2871 (2012).
10. M. E. J. Woolhouse, S. Gowtage-Sequeria, Host range and emerging and reemerging pathogens, *Emerg. Infect. Dis.* **11**, 1842–1847 (2005).
11. L. Brierley, A. B. Pedersen, M. E. J. Woolhouse, Tissue tropism and transmission ecology predict virulence of human RNA viruses, *PLoS Biol.* **17**, 1–18 (2019).
12. E. C. Holmes, *The Evolution and Emergence of RNA Viruses* (Oxford University Press, 2009).
13. J. Truscott, C. Fraser, S. Cauchemez, A. Meeyai, W. Hinsley, C. A. Donnelly, A. Ghani, N. Ferguson, Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza, *J. R. Soc. Interface* **9**, 304–312 (2012).
14. P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, M. A. Suchard, Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2, *PLoS Pathog.* **10** (2014), doi:10.1371/journal.ppat.1003932.
15. X. Wang, P. Wu, Y. Pei, T. K. Tsang, D. Gu, W. Wang, J. Zhang, P. W. Horby, T. M. Uyeki, B. J. Cowling, H. Yu, Assessment of human-to-human transmissibility of avian influenza A(H7N9) virus across 5 waves by analyzing clusters of case patients in Mainland China, 2013-2017, *Clin. Infect. Dis.* **68**, 623–631 (2019).
16. A. R. Fooks, F. Cliquet, S. Finke, C. Freuling, T. Hemachudha, R. S. Mani, T. Müller, S. Nadin-Davis, E. Picard-Meyer, H. Wilde, A. C. Banyard, Rabies, *Nat. Rev. Dis. Prim.* **3** (2017), doi:10.1038/nrdp.2017.91.
17. W. Rattanavipapong, M. Thavorncharoensap, S. Youngkong, A. J. Genuino, T. Anothaisintawee, U. Chaikledkaew, A. Meeyai, The impact of transmission dynamics of rabies control: Systematic review, *Vaccine* **37**, A154–A165 (2019).
18. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, MERS-CoV spillover at the camel-human interface, *Elife* **7**, 1–23 (2018).

19. C. A. Donnelly, M. C. Fisher, C. Fraser, A. C. Ghani, S. Riley, N. M. Ferguson, R. M. Anderson, Epidemiological and genetic analysis of severe acute respiratory syndrome, *Lancet Infect. Dis.* **4**, 672–683 (2004).
20. C. Beyrer, A pandemic anniversary: 40 years of HIV/AIDS, *Lancet* **397**, 2142–2143 (2021).
21. P. M. Sharp, B. H. Hahn, The evolution of HIV-1 and the origin of AIDS, *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2487–2494 (2010).
22. D. G. Streicker, A. S. Turmelle, M. J. Vonhof, I. V. Kuzmin, G. F. McCracken, C. E. Rupprecht, Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats, *Science (80-.).* **329**, 676–679 (2010).
23. M. J. Slusher, B. R. Wilcox, M. P. Lutrell, R. L. Poulson, J. D. Brown, M. J. Yabsley, D. E. Stallknecht, Are passerine birds reservoirs for influenza a viruses?, *J. Wildl. Dis.* **50**, 792–809 (2014).
24. C. R. Parrish, E. C. Holmes, D. M. Morens, E.-C. Park, D. S. Burke, C. H. Calisher, C. A. Laughlin, L. J. Saif, P. Daszak, Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases, *Microbiol. Mol. Biol. Rev.* **72**, 457–470 (2008).
25. M. Worobey, G. Z. Han, A. Rambaut, A synchronized global sweep of the internal genes of modern avian influenza virus, *Nature* **508**, 254–257 (2014).
26. T. T. Y. Lam, H. Zhu, Y. Guan, E. C. Holmes, Genomic Analysis of the Emergence, Evolution, and Spread of Human Respiratory RNA Viruses, *Annu. Rev. Genomics Hum. Genet.* **17**, 193–218 (2016).
27. A. E. Smith, A. Helenius, How Viruses Enter Animal Cells, *Science (80-.).* **304**, 237–242 (2004).
28. D. M. Knipe, P. Howley, *Fields Virology* W. & W. Lippincott, Ed. (ed. 6, 2013).
29. D. S. Dimitrov, Virus entry: Molecular mechanisms and biomedical applications, *Nat. Rev. Microbiol.* **2**, 109–122 (2004).
30. A. C. Lowen, Constraints, Drivers, and Implications of Influenza A Virus Reassortment, *Annu. Rev. Virol.* **4**, 105–121 (2017).
31. A. Heitmann, F. Gusmag, M. G. Rathjens, M. Maurer, K. Frankze, S. Schicht, S. Jansen, J. Schmidt-Chanasit, K. Jung, S. C. Becker, Mammals preferred: Reassortment of batai and bunyamwera orthobunyavirus occurs in mammalian but not insect cells, *Viruses* **13**, 1–14 (2021).
32. N. Conceição-Neto, J. R. Mesquita, M. Zeller, C. K. Yinda, F. Álvares, S. Roque, F. Petrucci-Fonseca, R. Godinho, E. Heylen, M. Van Ranst, J. Matthijnssens, Reassortment among picobirnaviruses found in wolves, *Arch. Virol.* **161**, 2859–2862 (2016).
33. T. Briese, C. H. Calisher, S. Higgs, Viruses of the family Bunyaviridae: Are all available isolates reassortants?, *Virology* **446**, 207–216 (2013).
34. S. M. McDonald, J. Matthijnssens, J. K. McAllen, E. Hine, L. Overton, S. Wang, P. Lemey, M. Zeller, M. Van Ranst, D. J. Spiro, J. T. Patton, Evolutionary dynamics of human rotaviruses: Balancing reassortment with preferred genome constellations, *PLoS Pathog.* **5** (2009), doi:10.1371/journal.ppat.1000634.
35. M. D. Stenglein, E. R. Jacobson, L. W. Chang, C. Sanders, M. G. Hawkins, D. S. M. Guzman, T. Drazenovich, F. Dunker, E. K. Kamaka, D. Fisher, D. R. Reavill, L. F.

Meola, G. Levens, J. L. DeRisi, Widespread Recombination, Reassortment, and Transmission of Unbalanced Compound Viral Genotypes in Natural Arenavirus Infections, *PLoS Pathog.* **11**, 1–26 (2015).
36. C.-C. Hon, T.-Y. Lam, A. Drummond, A. Rambaut, Y.-F. Lee, C.-W. Yip, F. Zeng, P.-Y. Lam, P. T. W. Ng, F. C. C. Leung, Phylogenetic Analysis Reveals a Correlation between the Expansion of Very Virulent Infectious Bursal Disease Virus and Reassortment of Its Genome Segment B, *J. Virol.* **80**, 8503–8509 (2006).
37. A. C. Lowen, It's in the mix: Reassortment of segmented viral genomes, *PLoS Pathog.* **14**, 1–6 (2018).
38. C. L. Parks, R. A. Lerch, P. Walpita, H.-P. Wang, M. S. Sidhu, S. A. Udem, Analysis of the Noncoding Regions of Measles Virus Strains in the Edmonston Vaccine Lineage, *J. Virol.* **75**, 921–933 (2001).
39. N. Chirico, A. Vianelli, R. Belshaw, Why genes overlap in viruses, *Proc. R. Soc. B Biol. Sci.* **277**, 3809–3817 (2010).
40. W. C. Ng, R. Soto-Acosta, S. S. Bradrick, M. A. Garcia-Blanco, E. E. Ooi, The 5′ and 3′ untranslated regions of the flaviviral genome, *Viruses* **9**, 1–14 (2017).
41. J. S. Y. Ho, Z. Zhu, I. Marazzi, Unconventional viral gene expression mechanisms as therapeutic targets, *Nature* **593**, 362–371 (2021).
42. N. H. L. Leung, Transmissibility and transmission of respiratory viruses, *Nat. Rev. Microbiol.* **19**, 528–545 (2021).
43. A. M. Gall, B. J. Mariñas, Y. Lu, J. L. Shisler, Waterborne Viruses: A Barrier to Safe Drinking Water, *PLoS Pathog.* **11**, 1–7 (2015).
44. B. M. Althouse, K. A. Hanley, The tortoise or the hare? Impacts of within-host dynamics on transmission success of arthropod-borne viruses, *Philos. Trans. R. Soc. B Biol. Sci.* **370** (2015), doi:10.1098/rstb.2014.0299.
45. M. Wille, E. C. Holmes, The ecology and evolution of influenza viruses, *Cold Spring Harb. Perspect. Med.* **10**, 1–19 (2020).
46. F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, R. A. Neher, Population genomics of intrapatient HIV-1 evolution, *Elife* **4**, 1–26 (2015).
47. A. M. Messenger, A. N. Barnes, G. C. Gray, Reverse zoonotic disease transmission (Zooanthroponosis): A systematic review of seldom-documented human biological threats to animals, *PLoS One* **9**, 1–9 (2014).
48. N. C. Grassly, C. Fraser, Mathematical models of infectious disease transmission, *Nat. Rev. Microbiol.* **6**, 477–487 (2008).
49. D. W. Redding, R. Gibb, C. C. Dan-Nwafor, E. A. Ilori, R. U. Yashe, S. H. Oladele, M. O. Amedu, A. Iniobong, L. A. Attfield, C. A. Donnelly, I. Abubakar, K. E. Jones, C. Ihekweazu, Geographical drivers and climate-linked dynamics of Lassa fever in Nigeria, *Nat. Commun.* **12**, 1–10 (2021).
50. E. Fichet-Calvet, E. Lecompte, L. Koivogui, B. Soropogui, A. Doré, F. Kourouma, O. Sylla, S. Daffis, K. Koulémou, J. Ter Meulen, Fluctuation of abundance and Lassa virus prevalence in Mastomys natalensis in Guinea, West Africa, *Vector-Borne Zoonotic Dis.* **7**, 119–128 (2007).
51. O. J. Brady, N. Golding, D. M. Pigott, M. U. G. Kraemer, J. P. Messina, R. C. R. Jr, T. W. Scott, D. L. Smith, P. W. Gething, S. I. Hay, Global temperature constraints

on Aedes aegypti and Ae . albopictus persistence and competence for dengue virus transmission, , 1–17 (2014).
52. M. U. G. Kraemer, M. E. Sinka, K. A. Duda, A. Q. N. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. Van Bortel, G. Hendrickx, F. Schaffner, I. R. Elyazar, H. J. Teng, O. J. Brady, J. P. Messina, D. M. Pigott, T. W. Scott, D. L. Smith, G. R. William Wint, N. Golding, S. I. Hay, The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus, *Elife* **4**, 1–18 (2015).
53. A. Chandrasekaran, A. Srinivasan, R. Raman, K. Viswanathan, S. Raguram, T. M. Tumpey, V. Sasisekharan, R. Sasisekharan, Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin, *Nat. Biotechnol.* **26**, 107–113 (2008).
54. T. Kuiken, E. C. Holmes, J. McCauley, G. F. Rimmelzwaan, C. S. Williams, B. T. Grenfell, Host species barriers to influenza virus infections, *Science (80-.).* **312**, 394–397 (2006).
55. K. Shinya, M. Ebina, S. Yamada, M. Ono, N. Kasai, Y. Kawaoka, Influenza virus receptors in the human airway, *Nature* **440**, 435–436 (2006).
56. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, *Nature* **581**, 215–220 (2020).
57. W. Li, M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough, H. Choe, M. Farzan, Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus, *Nature* **426**, 450–454 (2003).
58. J. Damas, G. M. Hughes, K. C. Keough, C. A. Painter, N. S. Persky, M. Corbo, M. Hiller, K. P. Koepfli, A. R. Pfenning, H. Zhao, D. P. Genereux, R. Swofford, K. S. Pollard, O. A. Ryder, M. T. Nweeia, K. Lindblad-Toh, E. C. Teeling, E. K. Karlsson, H. A. Lewin, Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates, *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22311–22322 (2020).
59. L. Lu, R. S. Sikkema, F. C. Velkers, D. F. Nieuwenhuijse, E. A. J. Fischer, P. A. Meijer, N. Bouwmeester-Vincken, A. Rietveld, M. C. A. Wegdam-Blans, P. Tolsma, M. Koppelman, L. A. M. Smit, R. W. Hakze-van der Honing, W. H. M. van der Poel, A. N. van der Spek, M. A. H. Spierenburg, R. J. Molenaar, J. de Rond, M. Augustijn, M. Woolhouse, J. A. Stegeman, S. Lycett, B. B. Oude Munnink, M. P. G. Koopmans, Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands, *Nat. Commun.* **12**, 1–12 (2021).
60. V. L. Hale, P. M. Dennis, D. S. McBride, J. M. Nolting, C. Madden, D. Huey, M. Ehrlich, J. Grieser, J. Winston, D. Lombardi, S. Gibson, L. Saif, M. L. Killian, K. Lantz, R. Tell, M. Torchetti, S. Robbe-Austerman, M. I. Nelson, S. A. Faith, A. S. Bowman, SARS-CoV-2 infection in free-ranging white-tailed deer, *Nature* (2021), doi:10.1038/s41586-021-04353-x.
61. C. R. Howard, N. F. Fletcher, Emerging virus diseases: Can we ever expect the unexpected?, *Emerg. Microbes Infect.* **1** (2012), doi:10.1038/emi.2012.47.
62. K. Gruber, Predicting zoonoses, *Nat. Ecol. Evol.* **1**, 1–4 (2017).
63. D. Carroll, B. Watson, E. Togami, P. Daszak, J. A. K. Mazet, C. J. Chrisman, E. M. Rubin, N. Wolfe, C. M. Morel, G. F. Gao, G. L. Burci, K. Fukuda, P. Auewarakul,

O. Tomori, Building a global atlas of zoonotic viruses, *Bull. World Health Organ.* **96**, 292–294 (2018).
64. B. Longdon, M. A. Brockhurst, C. A. Russell, J. J. Welch, F. M. Jiggins, The Evolution and Genetics of Virus Host Shifts, *PLoS Pathog.* **10** (2014), doi:10.1371/journal.ppat.1004395.
65. N. R. Faria, J. Quick, I. M. Claro, J. Thézé, J. G. De Jesus, M. Giovanetti, M. U. G. Kraemer, S. C. Hill, A. Black, A. C. Da Costa, L. C. Franco, S. P. Silva, C. H. Wu, J. Raghwani, S. Cauchemez, L. Du Plessis, M. P. Verotti, W. K. De Oliveira, E. H. Carmo, G. E. Coelho, A. C. F. S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. S. Nascimento, H. A. O. Monteiro, I. C. Siqueira, M. G. De Queiroz, T. R. De Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. De Lima, V. Nascimento, E. M. De Cerqueira, M. M. De Lima, D. L. Mascarenhas, J. P. M. Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. T. Nunes, E. C. Sabino, L. C. J. Alcantara, N. J. Loman, O. G. Pybus, Establishment and cryptic transmission of Zika virus in Brazil and the Americas, *Nature* **546**, 406–410 (2017).
66. R. C. Kading, A. C. Brault, J. D. Beckham, Global perspectives on arbovirus outbreaks: a 2020 snapshot, *Trop. Med. Infect. Dis.* **5**, 1–5 (2020).
67. F. Keesing, L. K. Belden, P. Daszak, A. Dobson, C. D. Harvell, R. D. Holt, P. Hudson, A. Jolles, K. E. Jones, C. E. Mitchell, S. S. Myers, T. Bogich, R. S. Ostfeld, Impacts of biodiversity on the emergence and transmission of infectious diseases, *Nature* **468**, 647–652 (2010).
68. B. A. Han, J. P. Schmidt, S. E. Bowden, J. M. Drake, Rodent reservoirs of future zoonotic diseases, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7039–7044 (2015).
69. S. L. Li, A. L. Acosta, S. C. Hill, O. J. Brady, M. A. B. de Almeida, J. da C. Cardoso, A. Hamlet, L. F. Mucci, J. Telles de Deus, F. C. M. Iani, N. S. Alexander, G. R. W. Wint, O. G. Pybus, M. U. G. Kraemer, N. R. Faria, J. P. Messina, Mapping environmental suitability of Haemagogus and Sabethes spp. mosquitoes to understand sylvatic transmission risk of yellow fever virus in Brazil, *PLoS Negl. Trop. Dis.* **16**, e0010019 (2022).
70. E. C. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. St. George, B. T. Grenfell, S. L. Salzberg, C. M. Fraser, D. J. Lipman, J. K. Taubenberger, Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses, *PLoS Biol.* **3**, 1579–1589 (2005).
71. W. Reed, J. Carroll, Yellow Fever : 100 Years of Discovery The Etiology of Yellow Fever : An Additional Note, *JAMA Class.* **36**, 1–3 (2008).
72. S. C. Weaver, Arrival of Chikungunya Virus in the New World: Prospects for Spread and Impact on Public Health, *PLoS Negl. Trop. Dis.* **8**, 6–9 (2014).
73. N. R. Faria, R. Do Socorro Da Silva Azevedo, M. U. G. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Thézé, M. B. Bonsall, T. A. Bowden, I. Rissanen, I. M. Rocco, J. S. Nogueira, A. Y. Maeda, F. G. Da Silva Vasami, F. L. De Lima Macedo, A. Suzuki,

S. G. Rodrigues, A. C. R. Cruz, B. T. Nunes, D. B. De Almeida Medeiros, D. S. G. Rodrigues, A. L. N. Queiroz, E. V. P. Da Silva, D. F. Henriques, E. S. T. Da Rosa, C. S. De Oliveira, L. C. Martins, H. B. Vasconcelos, L. M. N. Casseb, D. De Brito Simith, J. P. Messina, L. Abade, J. Lourenço, L. C. Alcantara, M. M. De Lima, M. Giovanetti, S. I. Hay, R. S. De Oliveira, P. Da Silva Lemos, L. F. De Oliveira, C. P. S. De Lima, S. P. Da Silva, J. M. De Vasconcelos, L. Franco, J. F. Cardoso, J. L. Da Silva Gonçalves Vianez-Júnior, D. Mir, G. Bello, E. Delatorre, K. Khan, M. Creatore, G. E. Coelho, W. K. De Oliveira, R. Tesh, O. G. Pybus, M. R. T. Nunes, P. F. C. Vasconcelos, Zika virus in the Americas: Early epidemiological and genetic findings, *Science (80-.).* **352**, 345–349 (2016).
74. N. S. D. Sahadeo, O. M. Allicock, P. M. De Salazar, A. J. Auguste, S. Widen, B. Olowokure, C. Gutierrez, A. M. Valadere, K. Polson-Edwards, S. C. Weaver, C. V. F. Carrington, Understanding the evolution and spread of chikungunya virus in the Americas using complete genome sequences, *Virus Evol.* **3**, 1–10 (2017).
75. S. E. Ronca, J. C. Ruff, K. O. Murray, A 20-year historical review of west nile virus since its initial emergence in north america: Has west nile virus become a neglected tropical disease?, *PLoS Negl. Trop. Dis.* **15**, 1–20 (2021).
76. E. C. Holmes, S. S. Twiddy, The origin, emergence and evolutionary genetics of dengue virus, *Infect. Genet. Evol.* **3**, 19–28 (2003).
77. C. Huang, O. A. Kolokoltsova, N. E. Yun, A. V. Seregin, S. Ronca, T. Koma, S. Paessler, Highly Pathogenic New World and Old World Human Arenaviruses Induce Distinct Interferon Responses in Human Cells, *J. Virol.* **89**, 7079–7088 (2015).
78. B. Hjelle, F. Torres-Pérez, Hantaviruses in the Americas and their role as emerging pathogens, *Viruses* **2**, 2559–2586 (2010).
79. C. Firth, R. Tokarz, D. B. Simith, M. R. T. Nunes, M. Bhat, E. S. T. Rosa, D. B. A. Medeiros, G. Palacios, P. F. C. Vasconcelos, W. I. Lipkin, Diversity and Distribution of Hantaviruses in South America, *J. Virol.* **86**, 13756–13766 (2012).
80. G. F. Albery, E. A. Eskew, N. Ross, K. J. Olival, Predicting the global mammalian viral sharing network using phylogeography, *Nat. Commun.* **11**, 1–9 (2020).
81. G. F. Albery, D. J. Becker, L. Brierley, C. E. Brook, R. C. Christofferson, L. E. Cohen, T. A. Dallas, E. A. Eskew, A. Fagre, M. J. Farrell, E. Glennon, S. Guth, M. B. Joseph, N. Mollentze, B. A. Neely, T. Poisot, A. L. Rasmussen, S. J. Ryan, S. Seifert, A. R. Sjodin, E. M. Sorrell, C. J. Carlson, The science of the host–virus network, *Nat. Microbiol.* **6**, 1483–1492 (2021).
82. F. Keesing, R. S. Ostfeld, Impacts of biodiversity and biodiversity loss on zoonotic diseases, *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–8 (2021).
83. O. Venter, E. W. Sanderson, A. Magrach, J. R. Allan, J. Beher, K. R. Jones, H. P. Possingham, W. F. Laurance, P. Wood, B. M. Fekete, M. A. Levy, J. E. M. Watson, Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation, *Nat. Commun.* **7**, 1–11 (2016).
84. M. R. T. Nunes, N. R. Faria, J. M. de Vasconcelos, N. Golding, M. U. G. Kraemer, L. F. de Oliveira, R. do S. da Silva Azevedo, D. E. A. da Silva, E. V. P. da Silva, S. P. da Silva, V. L. Carvalho, G. E. Coelho, A. C. R. Cruz, S. G. Rodrigues, J. L. da Silva Gonçalves Vianez, B. T. D. Nunes, J. F. Cardoso, R. B. Tesh, S. I. Hay, O. G. Pybus,

P. F. da Costa Vasconcelos, Emergence and potential for spread of Chikungunya virus in Brazil, *BMC Med.* **13** (2015), doi:10.1186/s12916-015-0348-x.
85. S. C. Hill, J. Vasconcelos, Z. Neto, D. Jandondo, L. Zé-Zé, R. S. Aguiar, J. Xavier, J. Thézé, M. Mirandela, A. L. Micolo Cândido, F. Vaz, C. dos S. Sebastião, C. H. Wu, M. U. G. Kraemer, A. Melo, B. L. F. Schamber-Reis, G. S. de Azevedo, A. Tanuri, L. M. Higa, C. Clemente, S. P. da Silva, D. da Silva Candido, I. M. Claro, D. Quibuco, C. Domingos, B. Pocongo, A. G. Watts, K. Khan, L. C. J. Alcantara, E. C. Sabino, E. Lackritz, O. G. Pybus, M. J. Alves, J. Afonso, N. R. Faria, Emergence of the Asian lineage of Zika virus in Angola: an outbreak investigation, *Lancet Infect. Dis.* **19**, 1138–1147 (2019).
86. I. I. Bogoch, O. J. Brady, M. U. G. Kraemer, M. German, M. I. Creatore, S. Brent, A. G. Watts, S. I. Hay, M. A. Kulkarni, J. S. Brownstein, K. Khan, Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study, *Lancet Infect. Dis.* **16**, 1237–1245 (2016).
87. O. Faye, M. de Lourdes Monteiro, B. Vrancken, M. Prot, S. Lequime, M. Diarra, O. Ndiaye, T. Valdez, S. Tavarez, J. Ramos, S. da Veiga Leal, C. Pires, A. Moreira, M. F. Tavares, L. Fernandes, J. N. Barreto, M. do Céu Teixeira, M. da Luz de Lima Mendonça, C. C. da Silva Leite Gomes, M. S. Castellon, L. Ma, F. Lemoine, F. Gámbaro-Roglia, D. Delaune, G. Fall, I. S. Fall, M. Diop, A. Sakuntabhai, C. Loucoubar, P. Lemey, E. C. Holmes, O. Faye, A. A. Sall, E. Simon-Loriere, Genomic epidemiology of 2015-2016 zika virus outbreak in Cape Verde, *Emerg. Infect. Dis.* **26**, 1084–1090 (2020).
88. R. S. Nasci, Movement of Chikungunya virus into the Western Hemisphere, *Emerg. Infect. Dis.* **20**, 1394–1395 (2014).
89. C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, et al, The Global Circulation of Seasonal Influenza A (H3N2) Viruses, *Science (80-.).* **320**, 340–347 (2008).
90. Y. C. F. Su, J. Bahl, U. Joseph, K. M. Butt, H. A. Peck, E. S. C. Koay, L. L. E. Oon, I. G. Barr, D. Vijaykrishna, G. J. D. Smith, Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection, *Nat. Commun.* **6** (2015), doi:10.1038/ncomms8952.
91. P. S. Born, M. M. Siqueira, N. R. Faria, P. C. Resende, F. C. Motta, G. Bello, Phylodynamics of influenza A(H3N2) in South America, 1999-2012, *Infect. Genet. Evol.* **43**, 312–320 (2016).
92. J. Mena, R. Tapia, C. Verdugo, L. Avendaño, P. Parra-Castro, R. A. Medina, G. Barriga, V. Neira, Circulation patterns of human seasonal Influenza A viruses in Chile before H1N1pdm09 pandemic, *Sci. Rep.* **11**, 1–9 (2021).
93. R. Palekar, A. Rodriguez, C. Avila, G. Barrera, M. Barrera, H. Brenes, A. Bruno, N. El Omeiri, R. Fasce, W. Ferreira de Almeida, D. Franco, M. Huaringa, J. Lara, R. Loayza, I. Lopez-Martinez, T. Maria de Paiva, J. Medina, J. Ojeda, A. M. Ropero, V. Sotomayor, C. Vazquez, M. Von Horoch, Patterns of influenza B circulation in Latin America and the Caribbean, 2010–2017, *PLoS One* **14**, 2010–2017 (2019).
94. T. S. Gritsun, E. A. Gould, Origin and evolution of flavivirus 5’ UTRs and panhandles: Trans-terminal duplications?, *Virology* **366**, 8–15 (2007).
95. G. R. Price, Selection and Covariance, *Nature* **227**, 520–521 (1970).

96. S. Duffy, L. A. Shackelton, E. C. Holmes, Rates of evolutionary change in viruses: Patterns and determinants, *Nat. Rev. Genet.* **9**, 267–276 (2008).
97. E. C. Holmes, The evolutionary genetics of emerging viruses, *Annu. Rev. Ecol. Evol. Syst.* **40**, 353–372 (2009).
98. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, 1992).
99. P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, K. H. Jacobsen, Complexity of the basic reproduction number (R0), *Emerg. Infect. Dis.* **25**, 1–4 (2019).
100. M. U. G. Kraemer, R. C. Reiner, O. J. Brady, J. P. Messina, M. Gilbert, D. M. Pigott, D. Yi, K. Johnson, L. Earl, L. B. Marczak, S. Shirude, N. Davis Weaver, D. Bisanzio, T. A. Perkins, S. Lai, X. Lu, P. Jones, G. E. Coelho, R. G. Carvalho, W. Van Bortel, C. Marsboom, G. Hendrickx, F. Schaffner, C. G. Moore, H. H. Nax, L. Bengtsson, E. Wetter, A. J. Tatem, J. S. Brownstein, D. L. Smith, L. Lambrechts, S. Cauchemez, C. Linard, N. R. Faria, O. G. Pybus, T. W. Scott, Q. Liu, H. Yu, G. R. W. Wint, S. I. Hay, N. Golding, Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus, *Nat. Microbiol.* **4**, 854–863 (2019).
101. L. J. S. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, *Infect. Dis. Model.* **2**, 128–142 (2017).
102. N. M. Ferguson, A. P. Galvanl, R. M. Bush, Ecological and immunological determinants of influenza evolution, *Nature* **422**, 428–433 (2003).
103. G. J. Abel, N. Sander, Quantifying Global International Migration Flows, *Science (80-.).* **343**, 1520–1523 (2014).
104. M. U. G. Kraemer, A. Sadilek, Q. Zhang, N. A. Marchal, G. Tuli, E. L. Cohn, Y. Hswen, T. A. Perkins, D. L. Smith, R. C. Reiner, J. S. Brownstein, Mapping global variation in human mobility, *Nat. Hum. Behav.* **4**, 800–810 (2020).
105. M. U. G. Kraemer, S. I. Hay, D. M. Pigott, D. L. Smith, G. R. W. Wint, N. Golding, Progress and Challenges in Infectious Disease Cartography, *Trends Parasitol.* **32**, 19–29 (2016).
106. J. M. Marshall, M. Touré, A. L. Ouédraogo, M. Ndhlovu, S. S. Kiware, A. Rezai, E. Nkhama, J. T. Griffin, T. D. Hollingsworth, S. Doumbia, N. J. Govella, N. M. Ferguson, A. C. Ghani, Key traveller groups of relevance to spatial malaria transmission: A survey of movement patterns in four sub-Saharan African countries, *Malar. J.* **15**, 1–12 (2016).
107. L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, F. Giannotti, Understanding the patterns of car travel, *Eur. Phys. J. Spec. Top.* **215**, 61–73 (2013).
108. A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, C. Mascolo, A tale of many cities: Universal patterns in human urban mobility, *PLoS One* **7** (2012), doi:10.1371/journal.pone.0037027.
109. M. Strohmeier, X. Olive, J. Lübbe, M. Schäfer, V. Lenders, Crowdsourced air traffic data from the OpenSky Network 2019-2020, *Earth Syst. Sci. Data* **13**, 357–366 (2021).
110. J. T. McCrone, A. S. Lauring, Genetic bottlenecks in intraspecies virus transmission, *Curr. Opin. Virol.* **28**, 20–25 (2018).
111. A. Rambaut, D. Posada, K. A. Crandall, E. C. Holmes, The causes and consequences of HIV evolution, *Nat. Rev. Genet.* **5**, 52–61 (2004).

112. J. O. Wertheim, A. M. Oster, B. Murrell, N. Saduvala, W. Heneine, W. M. Switzer, J. A. Johnson, Maintenance and reappearance of extremely divergent intra-host HIV-1 variants, *Virus Evol.* **4**, 1–9 (2018).
113. K. Theys, P. Libin, A. C. Pineda-Peña, A. Nowé, A. M. Vandamme, A. B. Abecasis, The impact of HIV-1 within-host evolution on transmission dynamics, *Curr. Opin. Virol.* **28**, 92–101 (2018).
114. R. A. Bull, F. Luciani, K. McElroy, S. Gaudieri, S. T. Pham, A. Chopra, B. Cameron, L. Maher, G. J. Dore, P. A. White, A. R. Lloyd, Sequential bottlenecks drive viral evolution in early acute hepatitis c virus infection, *PLoS Pathog.* **7** (2011), doi:10.1371/journal.ppat.1002243.
115. C. K. Y. Ho, J. Raghwani, S. Koekkoek, R. H. Liang, J. T. M. Van der Meer, M. Van Der Valk, M. De Jong, O. G. Pybus, J. Schinkel, R. Molenkamp, Characterization of Hepatitis C Virus (HCV) Envelope Diversification from Acute to Chronic Infection within a Sexually Transmitted HCV Cluster by Using Single-Molecule, Real-Time Sequencing, *J. Virol.* **91** (2017), doi:10.1128/jvi.02262-16.
116. E. C. Holmes, Patterns of Intra- and Interhost Nonsynonymous Variation Reveal Strong Purifying Selection in Dengue Virus, *J. Virol.* **77**, 11296–11298 (2003).
117. D. J. Park, G. Dudas, S. Wohl, A. Goba, S. L. M. Whitmer, K. G. Andersen, R. S. Sealfon, J. T. Ladner, J. R. Kugelman, C. B. Matranga, S. M. Winnicki, J. Qu, S. K. Gire, A. Gladden-Young, S. Jalloh, D. Nosamiefan, N. L. Yozwiak, L. M. Moses, P. P. Jiang, A. E. Lin, S. F. Schaffner, B. Bird, J. Towner, M. Mamoh, M. Gbakie, L. Kanneh, D. Kargbo, J. L. B. Massally, F. K. Kamara, E. Konuwa, J. Sellu, A. A. Jalloh, I. Mustapha, M. Foday, M. Yillah, B. R. Erickson, T. Sealy, D. Blau, C. Paddock, A. Brault, B. Amman, J. Basile, S. Bearden, J. Belser, E. Bergeron, S. Campbell, A. Chakrabarti, K. Dodd, M. Flint, A. Gibbons, C. Goodman, J. Klena, L. McMullan, L. Morgan, B. Russell, J. Salzer, A. Sanchez, D. Wang, I. Jungreis, C. Tomkins-Tinch, A. Kislyuk, M. F. Lin, S. Chapman, B. Macinnis, A. Matthews, J. Bochicchio, L. E. Hensley, J. H. Kuhn, C. Nusbaum, J. S. Schieffelin, B. W. Birren, M. Forget, S. T. Nichol, G. F. Palacios, D. Ndiaye, C. Happi, S. M. Gevao, M. A. Vandi, B. Kargbo, E. C. Holmes, T. Bedford, A. Gnirke, U. Ströher, A. Rambaut, R. F. Garry, P. C. Sabeti, Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone, *Cell* **161**, 1516–1526 (2015).
118. K. G. Andersen, B. J. Shapiro, C. B. Matranga, R. Sealfon, A. E. Lin, L. M. Moses, O. A. Folarin, A. Goba, I. Odia, P. E. Ehiane, M. Momoh, E. M. England, S. Winnicki, L. M. Branco, S. K. Gire, E. Phelan, R. Tariyal, R. Tewhey, O. Omoniwa, M. Fullah, R. Fonnie, M. Fonnie, L. Kanneh, S. Jalloh, M. Gbakie, S. Saffa, K. Karbo, A. D. Gladden, J. Qu, M. Stremlau, M. Nekoui, H. K. Finucane, S. Tabrizi, J. J. Vitti, B. Birren, M. Fitzgerald, C. McCowan, A. Ireland, A. M. Berlin, J. Bochicchio, B. Tazon-Vega, N. J. Lennon, E. M. Ryan, Z. Bjornson, D. A. Milner, A. K. Lukens, N. Broodie, M. Rowland, M. Heinrich, M. Akdag, J. S. Schieffelin, D. Levy, H. Akpan, D. G. Bausch, K. Rubins, J. B. McCormick, E. S. Lander, S. Günther, L. Hensley, S. Okogbenin, S. F. Schaffner, P. O. Okokhere, S. H. Khan, D. S. Grant, G. O. Akpede, D. A. Asogun, A. Gnirke, J. Z. Levin, C. T. Happi, R. F. Garry, P. C. Sabeti, Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus, *Cell* **162**, 738–750 (2015).

119. J. Hughes, R. C. Allen, M. Baguelin, K. Hampson, G. J. Baillie, D. Elton, J. R. Newton, P. Kellam, J. L. N. Wood, E. C. Holmes, P. R. Murcia, Transmission of Equine Influenza Virus during an Outbreak Is Characterized by Frequent Mixed Infections and Loose Transmission Bottlenecks, *PLoS Pathog.* **8**, 1–15 (2012).
120. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus, *J. Virol.* **91**, e00171-17 (2017).
121. C. J. R. Illingworth, A. Fischer, V. Mustonen, Identifying Selection in the Within-Host Evolution of Influenza Using Viral Sequence Data, *PLoS Comput. Biol.* **10** (2014), doi:10.1371/journal.pcbi.1003755.
122. D. K. Clarke, E. A. Duarte, A. Moya, S. F. Elena, E. Domingo, J. Holland, Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses, *J. Virol.* **67**, 222–228 (1993).
123. K. S. Xue, L. H. Moncla, T. Bedford, J. D. Bloom, Within-Host Evolution of Human Influenza Virus, *Trends Microbiol.* **26**, 781–793 (2018).
124. S. Alizon, F. Luciani, R. R. Regoes, Epidemiological and clinical consequences of within-host evolution, *Trends Microbiol.* **19**, 24–32 (2011).
125. E. Carrillo-Valenzo, R. Danis-Lozano, J. X. Velasco-Hernández, G. Sánchez-Burgos, C. Alpuche, I. López, C. Rosales, C. Baronti, X. de Lamballerie, E. C. Holmes, J. Ramos-Castañeda, Evolution of dengue virus in Mexico is characterized by frequent lineage replacement, *Arch. Virol.* **155**, 1401–1412 (2010).
126. C. Zhang, M. P. Mammen, P. Chinnawirotpisan, C. Klungthong, P. Rodpradit, P. Monkongdee, S. Nimmannitya, S. Kalayanarooj, E. C. Holmes, Clade Replacements in Dengue Virus Serotypes 1 and 3 Are Associated with Changing Serotype Prevalence, *J. Virol.* **79**, 15123–15130 (2005).
127. B. Adams, E. C. Holmes, C. Zhang, M. P. Mammen, S. Nimmannitya, S. Kalayanarooj, M. Boots, Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14234–14239 (2006).
128. K. Kupferschmidt, Evolving Threat, *Science (80-.).* **373**, 1–23 (2021).
129. Á. O'Toole, V. Hill, O. G. Pybus, A. Watts, I. I. Bogoch, K. Khan, J. P. Messina, B. C. G. Network, H. Tegally, R. R. Lessells, J. Giandhari, S. Pillay, K. A. Tumedi, G. Merhi, J. Koweyes, J. L. Geoghegan, J. De Ligt, X. Ren, M. Storey, N. E. Freed, C. Pattabiraman, P. Prasad, A. S. Desai, R. Vasanthapuram, T. F. Schulz, L. Leong, C. K. Lim, M. Chironna, D. Loconsole, E. Lasek-nesselquist, R. S. Sikkema, B. O. Munnink, Tracking the international spread of SARS-CoV-2 lineages B . 1 . 1 . 7 and B . 1 . 351 / 501Y-V2 [version 1 ; peer review : 3 approved] Network for Genomic Surveillance in South Africa (NGS-SA), Danish Covid-19 Genome Consortium (DCGC), Communicable, *Wellcome Open Res.* , 1–14 (2021).
130. J. T. McCrone, V. Hill, S. Bajaj, R. Evans Pena, B. C. Lambert, R. Inward, E. Al, Context-specific emergence and growth of the SARS-CoV-2 Delta variant John T. McCrone1,26, Verity Hill1,26, Sumali Bajaj2,26, Rosario Evans Pena2,26, Ben C. Lambert3, *medRxiv* **94**, 1–202 (2021).

131. R. Viana, S. Moy, D. G. Amoako, H. Tegally, C. Scheepers, E. Al, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa, *Nature* **Advance ar** (2022).
132. B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, E. C. Holmes, Unifying the Epidemiological and Evolutionary Dynamics of Pathogens, *Science (80-.).* **303**, 327–332 (2004).
133. L. Dyson, E. M. Hill, S. Moore, J. Curran-Sebastian, M. J. Tildesley, K. A. Lythgoe, T. House, L. Pellis, M. J. Keeling, Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics, *Nat. Commun.* **12** (2021), doi:10.1038/s41467-021-25915-7.
134. C. M. Saad-Roy, S. E. Morris, C. E. J. Metcalf, M. J. Mina, R. E. Baker, J. Farrar, E. C. Holmes, O. G. Pybus, A. L. Graham, S. A. Levin, B. T. Grenfell, C. E. Wagner, Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes, *Science (80-.).* **372**, 363–370 (2021).
135. L. Van Valen, A new evolutionary law, *Evol. Theory* **1**, 1–30 (1973).
136. M. A. Brockhurst, T. Chapman, K. C. King, J. E. Mank, S. Paterson, G. D. D. Hurst, Running with the Red Queen: The role of biotic conflicts in evolution, *Proc. R. Soc. B Biol. Sci.* **281** (2014), doi:10.1098/rspb.2014.1382.
137. K. Brunker, N. Mollentze, Rabies Virus, *Trends Microbiol.* **26**, 886–887 (2018).
138. D. G. Streicker, S. M. Altizer, A. Velasco-Villa, C. E. Rupprecht, Variable evolutionary routes to host establishment across repeated rabies virus host shifts among bats, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19715–19720 (2012).
139. N. R. Faria, M. U. G. Kraemer, S. C. Hill, J. Goes de Jesus, R. S. de Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C. H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolai, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa, S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. dos Santos, C. D. dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira, R. G. Abreu, M. H. O. Garcia, R. V. Alves, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, O. G. Pybus, Genomic and epidemiological monitoring of yellow fever virus transmission potential, *bioRxiv* **899**, 894–899 (2018).
140. A. B. Lacerda, L. del Castillo Saad, P. V. Ikefuti, A. Pinter, F. Chiaravalloti-Neto, Diffusion of sylvatic yellow fever in the state of São Paulo, Brazil, *Sci. Rep.* **11**, 1–11 (2021).
141. A. Moreira-Soto, M. C. Torres, M. C. Lima de Mendonça, M. A. Mares-Guia, C. D. dos Santos Rodrigues, A. A. Fabri, C. C. dos Santos, E. S. Machado Araújo, C. Fischer, R. M. Ribeiro Nogueira, C. Drosten, P. C. Sequeira, J. F. Drexler, A. M. Bispo de Filippis, Evidence for multiple sylvatic transmission cycles during the 2016–

2017 yellow fever virus outbreak, Brazil, *Clin. Microbiol. Infect.* **24**, 1019.e1-1019.e4 (2018).
142. A. Nasidi, T. P. Monath, K. Decock, R. Corderie, D. Olaleye, T. Ha, Yellow Fever Epidemic in Western, , 401–406 (1989).
143. M. Worobey, G. Z. Han, A. Rambaut, A synchronized global sweep of the internal genes of modern avian influenza virus, *Nature* **508**, 254–257 (2014).
144. R. Perez-Padilla, D. de la Rosa-Zamboni, S. Ponce de Leon, M. Hernandez, F. Quiñones-Falconi, E. Bautista, A. Ramirez-Venegas, J. Rojas-Serrano, C. E. Ormsby, A. Corrales, A. Higuera, E. Mondragon, J. A. Cordova-Villalobos, Pneumonia and Respiratory Failure from Swine-Origin Influenza A (H1N1) in Mexico, *N. Engl. J. Med.* **361**, 680–689 (2009).
145. C. F. Arias, M. Escalera-Zamudio, M. de los Dolores Soto-Del Río, A. Georgina Cobián-Güemes, P. Isa, S. López, Molecular Anatomy of 2009 Influenza virus A (H1N1), *Arch. Med. Res.* **40**, 643–654 (2009).
146. I. Mena, M. I. Nelson, F. Quezada-Monroy, J. Dutta, R. Cortes-Fernández, J. H. Lara-Puente, F. Castro-Peralta, L. F. Cunha, N. S. Trovão, B. Lozano-Dubernard, A. Rambaut, H. Van Bakel, A. García-Sastre, Origins of the 2009 H1N1 influenza pandemic in swine in Mexico, *Elife* **5**, 1–21 (2016).
147. T. J. Davies, A. B. Pedersen, Phylogeny and geography predict pathogen community similarity in wild primates and humans, *Proc. R. Soc. B Biol. Sci.* **275**, 1695–1701 (2008).
148. K. A. Tsetsarkin, R. Chen, M. B. Sherman, S. C. Weaver, Chikungunya virus: Evolution and genetic determinants of emergence, *Curr. Opin. Virol.* **1**, 310–317 (2011).
149. R. Chen, V. Puri, N. Fedorova, D. Lin, K. L. Hari, R. Jain, D. Rodas, R. S. Shabman, C. Weaver, Insights on the Global Expansion of Chikungunya Virus, *J. Virol.* **90**, 10600–10611 (2016).
150. R. S. Lanciotti, A. J. Lambert, M. Holodniy, S. Saavedra, C. Castillo, Phylogeny of Zika Virus in Western Hemisphere, 2015, *Emerg. Infect. Dis.* **22**, 933–935 (2016).
151. K. A. Tsetsarkin, D. L. Vanlandingham, C. E. McGee, S. Higgs, A single mutation in Chikungunya virus affects vector specificity and epidemic potential, *PLoS Pathog.* **3**, 1895–1906 (2007).
152. E. Lindh, C. Argentini, M. E. Remoli, C. Fortuna, G. Faggioni, E. Benedetti, A. Amendola, G. Marsili, F. Lista, G. Rezza, G. Venturi, The Italian 2017 outbreak chikungunya virus belongs to an emerging aedes albopictus-adapted virus cluster introduced from the Indian subcontinent, *Open Forum Infect. Dis.* **6** (2019), doi:10.1093/ofid/ofy321.
153. M. Escalera-Zamudio, M. Golden, B. Gutiérrez, J. Thézé, J. R. Keown, L. Carrique, T. A. Bowden, O. G. Pybus, Parallel evolution in the emergence of highly pathogenic avian influenza A viruses, *Nat. Commun.* **11**, 1–11 (2020).
154. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N. yuan Hsiao, S. Korsman, M. A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O.

Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa, *Nature* **592**, 438–443 (2021).
155. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. da S. Candido, S. Mishra, E. Al, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil, **372**, 815–821 (2021).
156. C. Padilla-Rojas, V. Jimenez-Vasquez, V. Hurtado, O. Mestanza, I. S. Molina, L. Barcena, S. Morales Ruiz, S. Acedo, W. Lizarraga, H. Bailon, O. Cáceres, M. Galarza, N. Rojas-Serrano, N. Vargas-Herrera, P. Lope-Pari, J. Huayra, L. Solari, Genomic analysis reveals a rapid spread and predominance of lambda (C.37) SARS-COV-2 lineage in Peru despite circulation of variants of concern, *J. Med. Virol.* **93**, 6845–6849 (2021).
157. P. E. Romero, G. Salvatierra, D. Cuicapuza, P. Marcos-carbajal, J. Huancachoque, L. Maturrano, P. Tsukayama, The Emergence of Sars-CoV-2 Variant Lambda (C.37) in South America, *Microbioogy Spectr.* **9**, e00789-21 (2021).
158. K. Laiton-Donato, C. Franco-Muñoz, D. A. Álvarez-Díaz, H. A. Ruiz-Moreno, J. A. Usme-Ciro, D. A. Prada, J. Reales-González, S. Corchuelo, M. T. Herrera-Sepúlveda, J. Naizaque, G. Santamaría, J. Rivera, P. Rojas, J. H. Ortiz, A. Cardona, D. Malo, F. Prieto-Alvarado, F. R. Gómez, M. Wiesner, M. L. O. Martínez, M. Mercado-Reyes, Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2, *Infect. Genet. Evol.* **95** (2021), doi:10.1016/j.meegid.2021.105038.
159. B. Gutierrez, S. Márquez, B. Prado-Vivar, M. Becerra-Wong, F. Zurita, E. Muñoz, J. J. Guadalupe, L. Patiño, A. Carrazco-Montalvo, J. C. Fernández-Cadena, D. Andrade-Molina, G. Morey-Leon, M. Martinez, F. Cardozo, M. E. Galeano, M. teresa Alvarez, O. Rollano, A. Vasquez, C. Salazar, I. Ferrés, P. Perbolianachis, M. Paz, A. Costábile, P. Moreno, G. Moratorio, G. Iraola, J. Coloma, P. Rojas-Silva, M. Grunauer, G. Trueba, V. Barragán, P. Cárdenas, Emergence of lineage B.1.621 in Latin America and the Caribbean*Virological.org* (2021).
160. S. Gupta, N. Ferguson, R. Anderson, Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents, *Science (80-.).* **280**, 912–915 (1998).
161. A. Moya, E. C. Holmes, F. González-Candelas, The population genetics and evolutionary epidemiology of RNA viruses, *Nat. Rev. Microbiol.* **2**, 279–288 (2004).
162. M. Ghafari, P. Simmonds, O. G. Pybus, A. Katzourakis, A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses, *Curr. Biol.* **31**, 4689-4696.e5 (2021).
163. M. Ghafari, L. du Plessis, J. Raghwani, S. Bhatt, B. Xu, O. G. Pybus, A. Katzourakis, Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza, *medRxiv* , 2021.07.27.21261148 (2021).
164. G. Loewenthal, D. Rapoport, O. Avram, A. Moshe, E. Wygoda, A. Itzkovitch, O. Israeli, D. Azouri, R. A. Cartwright, I. Mayrose, T. Pupko, A Probabilistic Model for Indel Evolution: Differentiating Insertions from Deletions, *Mol. Biol. Evol.* **38**, 5769–5781 (2021).

165. R. A. Cartwright, Problems and solutions for estimating indel rates and length distributions, *Mol. Biol. Evol.* **26**, 473–480 (2009).
166. B. Meng, S. A. Kemp, G. Papa, R. Datir, I. A. T. M. Ferreira, S. Marelli, W. T. Harvey, S. Lytras, A. Mohamed, G. Gallo, N. Thakur, D. A. Collier, P. Mlcochova, S. C. Robson, N. J. Loman, T. R. Connor, T. Golubchik, R. T. Martinez Nunez, C. Ludden, S. Corden, I. Johnston, D. Bonsall, C. P. Smith, A. R. Awan, G. Bucca, M. E. Torok, K. Saeed, J. A. Prieto, D. K. Jackson, W. L. Hamilton, L. B. Snell, C. Moore, E. M. Harrison, S. Goncalves, D. J. Fairley, M. W. Loose, J. Watkins, R. Livett, S. Moses, R. Amato, S. Nicholls, M. Bull, D. L. Smith, J. Barrett, D. M. Aanensen, M. D. Curran, S. Parmar, D. Aggarwal, J. G. Shepherd, M. D. Parker, S. Glaysher, M. Bashton, A. P. Underwood, N. Pacchiarini, K. F. Loveson, K. E. Templeton, C. F. Langford, J. Sillitoe, T. I. de Silva, D. Wang, D. Kwiatkowski, A. Rambaut, J. O'Grady, S. Cottrell, M. T. G. Holden, E. C. Thomson, H. Osman, M. Andersson, A. J. Chauhan, M. O. Hassan-Ibrahim, M. Lawniczak, A. Alderton, M. Chand, C. Constantinidou, M. Unnikrishnan, A. C. Darby, J. A. Hiscox, S. Paterson, I. Martincorena, E. M. Volz, A. J. Page, O. G. Pybus, A. R. Bassett, C. V. Ariani, M. H. S. Chapman, K. K. Li, R. N. Shah, N. G. Jesudason, Y. Taha, M. P. McHugh, R. Dewar, A. S. Jahun, C. McMurray, S. Pandey, J. P. McKenna, A. Nelson, G. R. Young, C. M. McCann, S. Elliott, H. Lowe, B. Temperton, S. Roy, A. Price, S. Rey, M. Wyles, S. Rooke, S. Shaaban, M. de Cesare, L. Letchford, S. Silveira, E. Pelosi, E. Wilson-Davies, M. Hosmillo, Á. O'Toole, A. R. Hesketh, R. Stark, L. du Plessis, C. Ruis, H. Adams, Y. Bourgeois, S. L. Michell, D. Gramatopoulos, J. Edgeworth, J. Breuer, J. A. Todd, C. Fraser, D. Buck, M. John, G. L. Kay, S. Palmer, S. J. Peacock, D. Heyburn, D. Weldon, E. Robinson, A. McNally, P. Muir, I. B. Vipond, J. Boyes, V. Sivaprakasam, T. Salluja, S. Dervisevic, E. J. Meader, N. R. Park, K. Oliver, A. R. Jeffries, S. Ott, A. da Silva Filipe, D. A. Simpson, C. Williams, J. A. H. Masoli, B. A. Knight, C. R. Jones, C. Koshy, A. Ash, A. Casey, A. Bosworth, L. Ratcliffe, L. Xu-McCrae, H. M. Pymont, S. Hutchings, L. Berry, K. Jones, F. Halstead, T. Davis, C. Holmes, M. Iturriza-Gomara, A. O. Lucaci, P. A. Randell, A. Cox, P. Madona, K. A. Harris, J. R. Brown, T. W. Mahungu, D. Irish-Tavares, T. Haque, J. Hart, E. Witele, M. L. Fenton, S. Liggett, C. Graham, E. Swindells, J. Collins, G. Eltringham, S. Campbell, P. C. McClure, G. Clark, T. J. Sloan, C. Jones, J. Lynch, B. Warne, S. Leonard, J. Durham, T. Williams, S. T. Haldenby, N. Storey, N. F. Alikhan, N. Holmes, C. Moore, M. Carlile, M. Perry, N. Craine, R. A. Lyons, A. H. Beckett, S. Goudarzi, C. Fearn, K. Cook, H. Dent, H. Paul, R. Davies, B. Blane, S. T. Girgis, M. A. Beale, K. L. Bellis, M. J. Dorman, E. Drury, L. Kane, S. Kay, S. McGuigan, R. Nelson, L. Prestwood, S. Rajatileka, R. Batra, R. J. Williams, M. Kristiansen, A. Green, A. Justice, A. I. K. Mahanama, B. Samaraweera, N. F. Hadjirin, J. Quick, R. Poplawski, L. M. Kermack, N. Reynolds, G. Hall, Y. Chaudhry, M. L. Pinckert, I. Georgana, R. J. Moll, A. Thornton, R. Myers, J. Stockton, C. A. Williams, W. C. Yew, A. J. Trotter, A. Trebes, G. MacIntyre-Cockett, A. Birchley, A. Adams, A. Plimmer, B. Gatica-Wilcox, C. McKerr, E. Hilvers, H. Jones, H. Asad, J. Coombes, J. M. Evans, L. Fina, L. Gilbert, L. Graham, M. Cronin, S. Kumziene-Summerhayes, S. Taylor, S. Jones, D. C. Groves, P. Zhang, M. Gallis, S. F. Louka, I. Starinskij, C. Jackson, M. Gourtovaia, G. Tonkin-Hill, K. Lewis, J. M. Tovar-Corona, K. James, L. Baxter, M. T.

Alam, R. J. Orton, J. Hughes, S. Vattipally, M. Ragonnet-Cronin, F. F. Nascimento, D. Jorgensen, O. Boyd, L. Geidelberg, A. E. Zarebski, J. Raghwani, M. U. G. Kraemer, J. Southgate, B. B. Lindsey, T. M. Freeman, J. P. Keatley, J. B. Singer, L. de Oliveira Martins, C. A. Yeats, K. Abudahab, B. E. W. Taylor, M. Menegazzo, J. Danesh, W. Hogsden, S. Eldirdiri, A. Kenyon, J. Mason, T. I. Robinson, A. Holmes, J. Price, J. A. Hartley, T. Curran, A. E. Mather, G. Shankar, R. Jones, R. Howe, S. Morgan, E. Wastenge, M. R. Chapman, S. Mookerjee, R. Stanley, W. Smith, T. Peto, D. Eyre, D. Crook, G. Vernet, C. Kitchen, H. Gulliver, I. Merrick, M. Guest, R. Munn, D. T. Bradley, T. Wyatt, C. Beaver, L. Foulser, S. Palmer, C. M. Churcher, E. Brooks, K. S. Smith, K. Galai, G. M. McManus, F. Bolt, F. Coll, L. Meadows, S. W. Attwood, A. Davies, E. De Lacy, F. Downing, S. Edwards, G. P. Scarlett, S. Jeremiah, N. Smith, D. Leek, S. Sridhar, S. Forrest, C. Cormie, H. K. Gill, J. Dias, E. E. Higginson, M. Maes, J. Young, M. Wantoch, D. Jamrozy, S. Lo, M. Patel, V. Hill, C. M. Bewshea, S. Ellard, C. Auckland, I. Harrison, C. Bishop, V. Chalker, A. Richter, A. Beggs, A. Best, B. Percival, J. Mirza, O. Megram, M. Mayhew, L. Crawford, F. Ashcroft, E. Moles-Garcia, N. Cumley, R. Hopes, P. Asamaphan, M. O. Niebel, R. N. Gunson, A. Bradley, A. Maclean, G. Mollett, R. Blacow, P. Bird, T. Helmer, K. Fallon, J. Tang, A. D. Hale, L. R. Macfarlane-Smith, K. L. Harper, H. Carden, N. W. Machin, K. A. Jackson, S. S. Y. Ahmad, R. P. George, L. Turtle, E. O'Toole, J. Watts, C. Breen, A. Cowell, A. Alcolea-Medina, T. Charalampous, A. Patel, L. J. Levett, J. Heaney, A. Rowan, G. P. Taylor, D. Shah, L. Atkinson, J. C. D. Lee, A. P. Westhorpe, R. Jannoo, H. L. Lowe, A. Karamani, L. Ensell, W. Chatterton, M. Pusok, A. Dadrah, A. Symmonds, G. Sluga, Z. Molnar, P. Baker, S. Bonner, S. Essex, E. Barton, D. Padgett, G. Scott, J. Greenaway, B. A. I. Payne, S. Burton-Fanning, S. Waugh, V. Raviprakash, N. Sheriff, V. Blakey, L. A. Williams, J. Moore, S. Stonehouse, L. Smith, R. K. Davidson, L. Bedford, L. Coupland, V. Wright, J. G. Chappell, T. Tsoleridis, J. Ball, M. Khakh, V. M. Fleming, M. M. Lister, H. C. Howson-Wells, L. Berry, T. Boswell, A. Joseph, I. Willingham, N. Duckworth, S. Walsh, E. Wise, N. Moore, M. Mori, N. Cortes, S. Kidd, R. Williams, L. Gifford, K. Bicknell, S. Wyllie, A. Lloyd, R. Impey, C. S. Malone, B. J. Cogger, N. Levene, L. Monaghan, A. J. Keeley, D. G. Partridge, M. Raza, C. Evans, K. Johnson, E. Betteridge, B. W. Farr, S. Goodwin, M. A. Quail, C. Scott, L. Shirley, S. A. J. Thurston, D. Rajan, I. F. Bronner, L. Aigrain, N. M. Redshaw, S. V. Lensing, S. McCarthy, A. Makunin, C. E. Balcazar, M. D. Gallagher, K. A. Williamson, T. D. Stanton, M. L. Michelsen, J. Warwick-Dugdale, R. Manley, A. Farbos, J. W. Harrison, C. M. Sambles, D. J. Studholme, A. Lackenby, T. Mbisa, S. Platt, S. Miah, D. Bibby, C. Manso, J. Hubb, G. Dabrera, M. Ramsay, D. Bradshaw, U. Schaefer, N. Groves, E. Gallagher, D. Lee, D. Williams, N. Ellaby, H. Hartman, N. Manesis, V. Patel, J. Ledesma, K. A. Twohig, E. Allara, C. Pearson, J. K. J. Cheng, H. E. Bridgewater, L. R. Frost, G. Taylor-Joyce, P. E. Brown, L. Tong, A. Broos, D. Mair, J. Nichols, S. N. Carmichael, K. L. Smollett, K. Nomikou, E. Aranday-Cortes, N. Johnson, S. Nickbakhsh, E. E. Vamos, M. Hughes, L. Rainbow, R. Eccles, C. Nelson, M. Whitehead, R. Gregory, M. Gemmell, C. Wierzbicki, H. J. Webster, C. L. Fisher, A. W. Signell, G. Betancor, H. D. Wilson, G. Nebbia, F. Flaviani, A. C. Cerda, T. V. Merrill, R. E. Wilson, M. Cotic, N. Bayzid, T. Thompson, E. Acheson, S. Rushton, S. O'Brien, D. J. Baker, S. Rudder, A. Aydin, F. Sang, J. Debebe, S.

Francois, T. I. Vasylyeva, M. E. Zamudio, B. Gutierrez, A. Marchbank, J. Maksimovic, K. Spellman, K. McCluggage, M. Morgan, R. Beer, S. Afifi, T. Workman, W. Fuller, C. Bresner, A. Angyal, L. R. Green, P. J. Parsons, R. M. Tucker, R. Brown, M. Whiteley, J. Bonfield, C. Puethe, A. Whitwham, J. Liddle, W. Rowe, I. Siveroni, T. Le-Viet, A. Gaskin, R. Johnson, I. Abnizova, M. Ali, L. Allen, R. Anderson, C. Ariani, S. Austin-Guest, S. Bala, J. Barrett, A. Bassett, K. Battleday, J. Beal, M. Beale, S. Bellany, T. Bellerby, K. Bellis, D. Berger, M. Berriman, P. Bevan, S. Binley, J. Bishop, K. Blackburn, N. Boughton, S. Bowker, T. Brendler-Spaeth, I. Bronner, T. Brooklyn, S. K. Buddenborg, R. Bush, C. Caetano, A. Cagan, N. Carter, J. Cartwright, T. C. Monteiro, L. Chapman, T. J. Chillingworth, P. Clapham, R. Clark, A. Clarke, C. Clarke, D. Cole, E. Cook, M. Coppola, L. Cornell, C. Cornwell, C. Corton, A. Crackett, A. Cranage, H. Craven, S. Craw, M. Crawford, T. Cutts, M. Dabrowska, M. Davies, J. Dawson, C. Day, A. Densem, T. Dibling, C. Dockree, D. Dodd, S. Dogga, M. Dorman, G. Dougan, M. Dougherty, A. Dove, L. Drummond, M. Dudek, L. Durrant, E. Easthope, S. Eckert, P. Ellis, B. Farr, M. Fenton, M. Ferrero, N. Flack, H. Fordham, G. Forsythe, M. Francis, A. Fraser, A. Freeman, A. Galvin, M. Garcia-Casado, A. Gedny, S. Girgis, J. Glover, O. Gould, A. Gray, E. Gray, C. Griffiths, Y. Gu, F. Guerin, W. Hamilton, H. Hanks, E. Harrison, A. Harrott, E. Harry, J. Harvison, P. Heath, A. Hernandez-Koutoucheva, R. Hobbs, D. Holland, S. Holmes, G. Hornett, N. Hough, L. Huckle, L. Hughes-Hallet, A. Hunter, S. Inglis, S. Iqbal, A. Jackson, D. Jackson, C. J. Verdejo, M. Jones, K. Kallepally, K. Kay, J. Keatley, A. Keith, A. King, L. Kitchin, M. Kleanthous, M. Klimekova, P. Korlevic, K. Krasheninnkova, G. Lane, C. Langford, A. Laverack, K. Law, S. Lensing, A. Lewis-Wade, J. Liddle, Q. Lin, S. Lindsay, S. Linsdell, R. Long, J. Lovell, J. Lovell, J. Mack, M. Maddison, A. Makunin, I. Mamun, J. Mansfield, N. Marriott, M. Martin, M. Mayho, J. McClintock, S. McHugh, L. MapcMinn, C. Meadows, E. Mobley, R. Moll, M. Morra, L. Morrow, K. Murie, S. Nash, C. Nathwani, P. Naydenova, A. Neaverson, E. Nerou, J. Nicholson, T. Nimz, G. G. Noell, S. O'Meara, V. Ohan, C. Olney, D. Ormond, A. Oszlanczi, Y. F. Pang, B. Pardubska, N. Park, A. Parmar, G. Patel, M. Payne, S. Peacock, A. Petersen, D. Plowman, T. Preston, M. Quail, R. Rance, S. Rawlings, N. Redshaw, J. Reynolds, M. Reynolds, S. Rice, M. Richardson, C. Roberts, K. Robinson, M. Robinson, D. Robinson, H. Rogers, E. M. Rojo, D. Roopra, M. Rose, L. Rudd, R. Sadri, N. Salmon, D. Saul, F. Schwach, P. Seekings, A. Simms, M. Sinnott, S. Sivadasan, B. Siwek, D. Sizer, K. Skeldon, J. Skelton, J. Slater-Tunstill, L. Sloper, N. Smerdon, C. Smith, C. Smith, J. Smith, K. Smith, M. Smith, S. Smith, T. Smith, L. Sneade, C. D. Soria, C. Sousa, E. Souster, A. Sparkes, M. Spencer-Chapman, J. Squares, R. Stanley, C. Steed, T. Stickland, I. Still, M. Stratton, M. Strickland, A. Swann, A. Swiatkowska, N. Sycamore, E. Swift, E. Symons, S. Szluha, E. Taluy, N. Tao, K. Taylor, S. Taylor, S. Thompson, M. Thompson, M. Thomson, N. Thomson, S. Thurston, D. Toombs, B. Topping, J. Tovar-Corona, D. Ungureanu, J. Uphill, J. Urbanova, P. J. Van, V. Vancollie, P. Voak, D. Walker, M. Walker, M. Waller, G. Ward, C. Weatherhogg, N. Webb, A. Wells, E. Wells, L. Westwood, T. Whipp, T. Whiteley, G. Whitton, S. Widaa, M. Williams, M. Wilson, S. Wright, L. M. Duncan, A. M. Carabelli, J. C. Kenyon, A. M. Lever, A. De Marco, C. Saliba, K. Culap, E. Cameroni, N. J. Matheson, L. Piccoli, D. Corti, L. C. James, D. L. Robertson, D. Bailey, R. K. Gupta,

Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7, *Cell Rep.* **35** (2021), doi:10.1016/j.celrep.2021.109292.
167. K. R. McCarthy, L. J. Rennick, S. Nambulli, L. R. Robinson-McCarthy, W. G. Bain, G. Haidar, W. Paul Duprex, Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape, *Science (80-.).* **371**, 1139–1142 (2021).
168. N. Nao, J. Yamagishi, H. Miyamoto, M. Igarashi, R. Manzoor, A. Ohnuma, Y. Tsuda, W. Furuyama, A. Shigeno, M. Kajihara, N. Kishida, R. Yoshida, A. Takada, Genetic Predisposition To Acquire a Polybasic Cleavage Site for Highly, *MBio* **8**, e02298-16 (2017).
169. P. Simmonds, A. Tuplin, D. J. Evans, Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence., *RNA* **10**, 1337–1351 (2004).
170. M. Davis, S. M. Sagan, J. P. Pezacki, D. J. Evans, P. Simmonds, Bioinformatic and Physical Characterizations of Genome-Scale Ordered RNA Structure in Mammalian RNA Viruses, *J. Virol.* **82**, 11824–11836 (2008).
171. Y. Liu, Y. Zhang, M. Wang, A. Cheng, Q. Yang, Y. Wu, R. Jia, M. Liu, D. Zhu, S. Chen, S. Zhang, X. X. Zhao, J. Huang, S. Mao, X. Ou, Q. Gao, Y. Wang, Z. Xu, Z. Chen, L. Zhu, Q. Luo, Y. Liu, Y. Yu, L. Zhang, B. Tian, L. Pan, X. Chen, Structures and Functions of the 3′ Untranslated Regions of Positive-Sense Single-Stranded RNA Viruses Infecting Humans and Animals, *Front. Cell. Infect. Microbiol.* **10**, 1–16 (2020).
172. N. McFadden, A. Arias, I. Dry, D. Bailey, J. Witteveldt, D. J. Evans, I. Goodfellow, P. Simmonds, Influence of genome-scale RNA structure disruption on the replication of murine norovirus - Similar replication kinetics in cell culture but attenuation of viral fitness in vivo, *Nucleic Acids Res.* **41**, 6316–6331 (2013).
173. E. Simon-Loriere, E. C. Holmes, Why do RNA viruses recombine?, *Nat. Rev. Microbiol.* **9**, 617–626 (2011).
174. Y. Xiao, I. M. Rouzine, S. Bianco, A. Acevedo, E. F. Goldstein, M. Farkov, L. Brodsky, R. Andino, RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence, *Cell Host Microbe* **19**, 493–503 (2016).
175. H. Song, E. E. Giorgi, V. V. Ganusov, F. Cai, G. Athreya, H. Yoon, O. Carja, B. Hora, P. Hraber, E. Romero-Severson, C. Jiang, X. Li, S. Wang, H. Li, J. F. Salazar-Gonzalez, M. G. Salazar, N. Goonetilleke, B. F. Keele, D. C. Montefiori, M. S. Cohen, G. M. Shaw, B. H. Hahn, A. J. McMichael, B. F. Haynes, B. Korber, T. Bhattacharya, F. Gao, Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection, *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-04217-5.
176. K. B. Westgeest, C. A. Russell, X. Lin, M. I. J. Spronken, T. M. Bestebroer, J. Bahl, R. van Beek, E. Skepner, R. A. Halpin, J. C. de Jong, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, D. J. Smith, D. E. Wentworth, R. A. M. Fouchier, M. de Graaf, A. Garcia-Sastre, Genomewide Analysis of Reassortment and Evolution of Human Influenza A(H3N2) Viruses Circulating between 1968 and 2011, *J. Virol.* **88**, 2844–2857 (2014).
177. V. N. Petrova, C. A. Russell, The evolution of seasonal influenza viruses, *Nat. Rev. Microbiol.* **16**, 47–60 (2018).

178. J. S. M. Sabir, T. T. Y. Lam, M. M. M. Ahmed, L. Li, Y. Shen, S. E. M. Abo-Aba, M. I. Qureshi, M. Abu-Zeid, Y. Zhang, M. A. Khiyami, N. S. Alharbi, N. H. Hajrah, M. J. Sabir, M. H. Z. Mutwakil, S. A. Kabli, F. A. S. Alsulaimany, A. Y. Obaid, B. Zhou, D. K. Smith, E. C. Holmes, H. Zhu, Y. Guan, Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia, *Science (80-.).* **351**, 81–84 (2016).
179. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century, *Nat. Microbiol.* **4**, 10–19 (2019).
180. V. Hill, C. Ruis, S. Bajaj, O. G. Pybus, M. U. G. Kraemer, Progress and challenges in virus genomic epidemiology, *Trends Parasitol.* **37**, 1038–1049 (2021).
181. L. Du Plessis, T. Stadler, Getting to the root of epidemic spread with phylodynamic analysis of genomic data, *Trends Microbiol.* **23**, 383–386 (2015).
182. H. min Neoh, X. E. Tan, H. F. Sapri, T. L. Tan, Pulsed-field gel electrophoresis (PFGE): A review of the "gold standard" for bacteria typing and current alternatives, *Infect. Genet. Evol.* **74**, 103935 (2019).
183. M. D. Fallin, P. Duggal, T. H. Beaty, Genetic epidemiology and public health: The evolution from theory to technology, *Am. J. Epidemiol.* **183**, 387–393 (2016).
184. M. K. Grabowski, A. D. Redd, Molecular tools for studying HIV transmission in sexual networks, *Curr. Opin. HIV AIDS* **9**, 126–133 (2014).
185. A. J. Drummond, O. G. Pybus, A. Rambaut, R. Forsberg, A. G. Rodrigo, Measurably evolving populations, *Trends Ecol. Evol.* **18**, 481–488 (2003).
186. N. R. Pace, J. Sapp, N. Goldenfeld, Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1011–1018 (2012).
187. M. Haber, J. Velasco, in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (Metaphysics Research Lab, Stanford University, 2021).
188. J. Felsenstein, Confidence Limits on Phylogenies: an Approach Using the Bootstrap, *Evolution (N. Y).* **39**, 783–791 (1985).
189. M. Anisimova, O. Gascuel, Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative, *Syst. Biol.* **55**, 539–552 (2006).
190. P. Lemey, M. Salemi, A. M. Vandamme, *The Phylogenetic Handbook: A Practical Approach to Phylognetic Analysis and Hypothesis Testing* (Cambridge University Press, Cambridge, UK, Second Edi., 2009).
191. Z. Yang, B. Rannala, Molecular phylogenetics: Principles and practice, *Nat. Rev. Genet.* **13**, 303–314 (2012).
192. D. Posada, K. A. Crandall, MODELTEST: testing the model of DNA substitution, *Bioinformatics* **14**, 817–818 (1998).
193. J. Joyce, in *Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (Metaphysics Research Lab, Stanford University, 2003).
194. J. Heled, R. R. Bouckaert, Looking for trees in the forest: Summary tree from posterior samples, *BMC Evol. Biol.* **13** (2013), doi:10.1186/1471-2148-13-221.
195. L. Bromham, D. Penny, The modern molecular clock, *Nat. Rev. Genet.* **4**, 216–224 (2003).

196. E. Zuckerkandl, L. Pauling, Molecules as documents of history, *J. Theor. Biol.* **8**, 357–366 (1965).
197. S. Kumar, Molecular clocks: four decades of evolution., *Nat. Rev. Genet.* **6**, 654–662 (2005).
198. P. Aiewsakun, A. Katzourakis, Time-Dependent Rate Phenomenon in Viruses, *J. Virol.* **90**, 7184–7195 (2016).
199. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence, *PLoS Biol.* **4**, 699–710 (2006).
200. P. Aiewsakun, A. Katzourakis, Time dependency of foamy virus evolutionary rate estimates, *BMC Evol. Biol.* **15**, 1–15 (2015).
201. M. A. Suchard, A. J. Drummond, Bayesian random local clocks, or one rate to rule them all, *BMC Biol.* **8**, 114 (2010).
202. A. Rambaut, T. T. Lam, L. M. Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen), *Virus Evol.* **2**, 1–7 (2016).
203. S. Duchêne, D. Duchêne, E. C. Holmes, S. Y. W. Ho, The performance of the date-randomization test in phylogenetic analyses of time-structured virus data, *Mol. Biol. Evol.* **32**, 1895–1906 (2015).
204. A. Rambaut, Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies, *Bioinformatics* **16**, 395–399 (2000).
205. S. Hohna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, F. Ronquist, RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language, *Syst. Biol.* **65**, 726–736 (2016).
206. J. P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics* **17**, 754–755 (2001).
207. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, *Virus Evol.* **4**, 1–5 (2018).
208. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7., *Mol. Biol. Evol.* **29**, 1969–73 (2012).
209. R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. Du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C. H. Wu, D. Xie, C. Zhang, T. Stadler, A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis, *PLoS Comput. Biol.* **15**, 1–28 (2019).
210. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis, *Virus Evol.* **4**, 1–9 (2018).
211. E. W. Bloomquist, P. Lemey, M. A. Suchard, Three roads diverged? Routes to phylogeographic inference, *Trends Ecol. Evol.* **25**, 626–632 (2010).
212. R. Beard, D. Magee, M. A. Suchard, P. Lemey, M. Scotch, Generalized linear models for identifying predictors of the evolutionary diffusion of viruses., *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.* **2014**, 23–8 (2014).

213. T. G. Vaughan, D. Kühnert, A. Popinga, D. Welch, A. J. Drummond, Efficient Bayesian inference under the structured coalescent, *Bioinformatics* **30**, 2272–2279 (2014).
214. N. De Maio, C. H. Wu, K. M. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, *PLoS Genet.* **11**, 1–22 (2015).
215. D. Kühnert, T. Stadler, T. G. Vaughan, A. J. Drummond, Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data, *Mol. Biol. Evol.* **33**, 2102–2116 (2016).
216. M. Slatkin, W. P. Maddison, A cladistic measure of gene flow inferred from the phylogenies of alleles, *Genetics* **123**, 603–613 (1989).
217. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots, *PLoS Comput. Biol.* **5** (2009), doi:10.1371/journal.pcbi.1000520.
218. E. Paradis, S. R. Baillie, W. J. Sutherland, Modeling large-scale dispersal distances, *Ecol. Modell.* **151**, 279–292 (2002).
219. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time, *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
220. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution, *J. Math. Biol.* **56**, 391–412 (2008).
221. J. F. C. Kingman, The coalescent, *Stoch. Process. their Appl.* **13**, 235–248 (1982).
222. P. Donelly, S. Tavaré, Coalescents and Genealogical Structure Under Neutrality, *Annu. Rev. Genet.* **29**, 401–421 (1995).
223. O. G. Pybus, A. Rambaut, P. H. Harvey, An integrated framework for the inference of viral population history from reconstructed genealogies, *Genetics* **155**, 1429–1437 (2000).
224. A. Eriksson, B. Mehlig, M. Rafajlovic, S. Sagitov, The total branch length of sample genealogies in populations of variable size, *Genetics* **186**, 601–611 (2010).
225. T. Stadler, M. Steel, Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models, *J. Theor. Biol.* **297**, 33–40 (2012).
226. R. C. Griffiths, S. Tavaré, Sampling theory for neutral alleles in a varying environment, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **344**, 403–410 (1994).
227. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences, *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
228. S. Möller, L. du Plessis, T. Stadler, Impact of the tree prior on estimating clock rates during epidemic outbreaks, *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4200–4205 (2018).
229. O. G. Pybus, A. Rambaut, GENIE: Estimating demographic history from molecular phylogenies, *Bioinformatics* **18**, 1404–1405 (2002).
230. V. Hill, G. Baele, Bayesian Estimation of Past Population Dynamics in BEAST 1.10 Using the Skygrid Coalescent Model, *Mol. Biol. Evol.* **36**, 2620–2628 (2019).
231. K. Strimmer, O. G. Pybus, Exploring the demographic history of DNA sequences using the generalized skyline plot, *Mol. Biol. Evol.* **18**, 2298–2305 (2001).

232. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics, *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
233. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving bayesian population dynamics inference: A coalescent-based model for multiple loci, *Mol. Biol. Evol.* **30**, 713–724 (2013).
234. M. S. Gill, P. Lemey, S. N. Bennett, R. Biek, M. A. Suchard, Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates, *Syst. Biol.* **65**, 1041–1056 (2016).
235. T. Ohta, Slightly Deleterious Mutant Substitutions in Evolution, *Nature* **246**, 96–98 (1973).
236. T. Ohta, J. H. Gillespie, Development of Neutral and Nearly Neutral Theories, *Theor. Popul. Biol.* **49**, 128–142 (1996).
237. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations, *Nat. Rev. Genet.* **8**, 610–618 (2007).
238. A. Eyre-Walker, P. D. Keightley, Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change, *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
239. P. D. Keightley, A. Eyre-Walker, What can we learn about the distribution of fitness effects of new mutations from DNA sequence data?, *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 1187–1193 (2010).
240. S. Bhatt, A. Katzourakis, O. G. Pybus, Detecting natural selection in RNA virus populations using sequence summary statistics, *Infect. Genet. Evol.* **10**, 421–430 (2010).
241. S. L. Kosakovsky Pond, S. D. W. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies, *Bioinformatics* **21**, 676–679 (2005).
242. S. L. Kosakovsky Pond, A. F. Y. Poon, R. Velazquez, S. Weaver, N. L. Hepler, B. Murrell, S. D. Shank, B. R. Magalis, D. Bouvier, A. Nekrutenko, S. Wisotsky, S. J. Spielman, S. D. W. Frost, S. V. Muse, HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies, *Mol. Biol. Evol.* **37**, 295–299 (2020).
243. Z. Yang, R. Nielsent, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages, *Mol. Biol. Evol.* **19**, 908–917 (2002).
244. B. Murrell, S. Moola, A. Mabona, T. Weighill, D. Sheward, S. L. Kosakovsky Pond, K. Scheffler, FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection, *Mol. Biol. Evol.* **30**, 1196–1205 (2013).
245. S. L. Kosakovsky Pond, S. D. W. Frost, Not so different after all: A comparison of methods for detecting amino acid sites under selection, *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
246. M. D. Smith, J. O. Wertheim, S. Weaver, B. Murrell, K. Scheffler, S. L. Kosakovsky Pond, Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection, *Mol. Biol. Evol.* **32**, 1342–1353 (2015).

247. B. Murrell, J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S. L. Kosakovsky Pond, Detecting individual sites subject to episodic diversifying selection, *PLoS Genet.* **8** (2012), doi:10.1371/journal.pgen.1002764.
248. P. Lemey, V. N. Minin, F. Bielejec, S. L. Kosakovsky Pond, M. A. Suchard, A counting renaissance: Combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection, *Bioinformatics* **28**, 3248–3256 (2012).
249. M. Pérez-Losada, M. Arenas, J. C. Galán, F. Palero, F. González-Candelas, Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences, *Infect. Genet. Evol.* **30**, 296–307 (2015).
250. S. M. McDonald, M. I. Nelson, P. E. Turner, J. T. Patton, Reassortment in segmented RNA viruses: Mechanisms and outcomes, *Nat. Rev. Microbiol.* **14**, 448–460 (2016).
251. J. Steel, A. C. Lowen, in *Influenza Pathogenesis and Control - Volume 1*, R. W. Compans, M. B. A. Oldstone, Eds. (Springer, Cham, 2014), vol. Volume 1, pp. 377–401.
252. D. Martin, E. Rybicki, RDP: detection of recombination amongst aligned sequences. Bioinformatics, *Oxford Journals* **16**, 562–563 (2000).
253. D. Posada, Evaluation of methods for detecting recombination from DNA sequences: Empirical data, *Mol. Biol. Evol.* **19**, 708–717 (2002).
254. M. I. Nelson, C. Viboud, L. Simonsen, R. T. Bennett, S. B. Griesemer, K. St. George, J. Taylor, D. J. Spiro, N. A. Sengamalay, E. Ghedin, J. K. Taubenberger, E. C. Holmes, Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918, *PLoS Pathog.* **4** (2008), doi:10.1371/journal.ppat.1000012.
255. L. Lu, S. J. Lycett, A. J. Leigh Brown, Reassortment patterns of avian influenza virus internal segments among different subtypes, *BMC Evol. Biol.* **14**, 1–15 (2014).
256. G. Dudas, T. Bedford, S. Lycett, A. Rambaut, Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1-PB2-HA Gene Complex, *Mol. Biol. Evol.* **32**, 162–172 (2015).
257. M. O. Salminen, J. K. Carr, D. S. Burke, F. E. Mccutchan, Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning, *AIDS Res. Hum. Retroviruses* **11**, 1423–1425 (1995).
258. M. Worobey, E. C. Holmes, Evolutionary aspects of recombination in RNA viruses, *J. Gen. Virol.* **80**, 2535–2543 (1999).
259. S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, S. D. W. Frost, GARD: A genetic algorithm for recombination detection, *Bioinformatics* **22**, 3096–3098 (2006).
260. S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, S. D. W. Frost, Automated phylogenetic detection of recombination using a genetic algorithm, *Mol. Biol. Evol.* **23**, 1891–1901 (2006).
261. J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. J. Baum, B. Becker-Ziaja, J. P. Boettcher, M. Cabeza-Cabrerizo, Á. Camino-Sánchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. Garciá-Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallasch, L.

V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, T. Racine, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. J. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. D. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Wolfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keita, A. Rambaut, P. Formenty, S. Gunther, M. W. Carroll, Real-time, portable genome sequencing for Ebola surveillance, *Nature* **530**, 228–232 (2016).

262. N. R. Faria, E. C. Sabino, M. R. T. Nunes, L. C. J. Alcantara, N. J. Loman, O. G. Pybus, Mobile real-time surveillance of Zika virus in Brazil, *Genome Med.* **8**, 2–5 (2016).

263. L. W. Meredith, W. L. Hamilton, B. Warne, C. J. Houldcroft, M. Hosmillo, A. S. Jahun, M. D. Curran, S. Parmar, L. G. Caller, S. L. Caddy, F. A. Khokhar, A. Yakovleva, G. Hall, T. Feltwell, S. Forrest, S. Sridhar, M. P. Weekes, S. Baker, N. Brown, E. Moore, A. Popay, I. Roddick, M. Reacher, T. Gouliouris, S. J. Peacock, G. Dougan, M. E. Török, I. Goodfellow, Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study, *Lancet Infect. Dis.* **20**, 1263–1272 (2020).

264. K. Brunker, G. Jaswant, S. M. Thumbi, K. Lushasi, A. Lugelo, A. M. Czupryna, F. Ade, G. Wambura, V. Chuchu, R. Steenson, C. Ngeleja, C. Bautista, D. L. Manalo, M. R. R. Gomez, M. Y. J. V. Chu, M. E. Miranda, M. Kamat, K. Rysava, J. Espineda, E. A. V. Silo, A. M. Aringo, R. P. Bernales, F. F. Adonay, M. J. Tildesley, D. A. Marston, D. L. Jennings, A. R. Fooks, W. Zhu, L. W. Meredith, S. C. Hill, R. Poplawski, R. J. Gifford, J. B. Singer, M. Maturi, A. Mwatondo, R. Biek, K. Hampson, Rapid in-country sequencing of whole virus genomes to inform rabies elimination programmes [version 1; peer review: 3 approved], *Wellcome Open Res.* **5**, 1–31 (2020).

265. G. J. D. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt, J. S. M. Peiris, Y. Guan, A. Rambaut, Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic, *Nature* **459**, 1122–1125 (2009).

266. H. K. H. Luk, X. Li, J. Fung, S. K. P. Lau, P. C. Y. Woo, Molecular epidemiology, evolution and phylogeny of SARS coronavirus, *Infect. Genet. Evol.* **71**, 21–30 (2019).

267. B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, W. Liang, H. Zheng, K. Wan, Q. Liu, B. Cui, Y. Xu, E. Zhang, H. Wang, J. Ye, G. Li, M. Li, Z. Cui, X. Qi, K. Chen, L. Du, K. Gao, Y. Zhao, X. Zou, Y.-J. Feng, Y.-F. Gao, R. Hai, D. Yu, Y. Guan, J. Xu, Molecular Evolution Analysis and Geographic Investigation of Severe Acute Respiratory Syndrome Coronavirus-Like Virus in Palm Civets at an Animal Market and on Farms, *J. Virol.* **79**, 11892–11900 (2005).

268. Y. Guan, B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, P. H. Li, L. J. Zhang, Y. J. Guan, K. M. Butt, K. L. Wong, K. W. Chan, W. Lim, K. F. Shortridge, K. Y. Yuen, J. S. M. Peiris, L. L. M. Poon, Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China, *Science (80-.).* **302**, 276–278 (2003).
269. H. D. Song, C. C. Tu, G. W. Zhang, S. Y. Wang, K. Zheng, L. C. Lei, Q. X. Chen, Y. W. Gao, H. Q. Zhou, H. Xiang, H. J. Zheng, S. W. W. Chern, F. Cheng, C. M. Pan, H. Xuan, S. J. Chen, H. M. Luo, D. H. Zhou, Y. F. Liu, J. F. He, P. Z. Qin, L. H. Li, Y. Q. Ren, W. J. Liang, Y. D. Yu, L. Anderson, M. Wang, R. H. Xu, X. W. Wu, H. Y. Zheng, J. D. Chen, G. Liang, Y. Gao, M. Liao, L. Fang, L. Y. Jiang, H. Li, F. Chen, B. Di, L. J. He, J. Y. Lin, S. Tong, X. Kong, L. Du, P. Hao, H. Tang, A. Bernini, X. J. Yu, O. Spiga, Z. M. Guo, H. Y. Pan, W. Z. He, J. C. Manuguerra, A. Fontanet, A. Danchin, N. Niccolai, Y. X. Li, C. I. Wu, G. P. Zhao, Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2430–2435 (2005).
270. S. K. P. Lau, P. C. Y. Woo, K. S. M. Li, Y. Huang, H. W. Tsoi, B. H. L. Wong, S. S. Y. Wong, S. Y. Leung, K. H. Chan, K. Y. Yuen, Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14040–14045 (2005).
271. S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, P. C. Sabeti, Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 Outbreak, *Science (80-.).* **345**, 1369–1372 (2014).
272. D. J. Park, G. Dudas, S. Wohl, A. Goba, S. L. M. Whitmer, K. G. Andersen, R. S. Sealfon, J. T. Ladner, J. R. Kugelman, C. B. Matranga, S. M. Winnicki, J. Qu, S. K. Gire, A. Gladden-Young, S. Jalloh, D. Nosamiefan, N. L. Yozwiak, L. M. Moses, P. P. Jiang, A. E. Lin, S. F. Schaffner, B. Bird, J. Towner, M. Mamoh, M. Gbakie, L. Kanneh, D. Kargbo, J. L. B. Massally, F. K. Kamara, E. Konuwa, J. Sellu, A. A. Jalloh, I. Mustapha, M. Foday, M. Yillah, B. R. Erickson, T. Sealy, D. Blau, C. Paddock, A. Brault, B. Amman, J. Basile, S. Bearden, J. Belser, E. Bergeron, S. Campbell, A. Chakrabarti, K. Dodd, M. Flint, A. Gibbons, C. Goodman, J. Klena, L. McMullan, L. Morgan, B. Russell, J. Salzer, A. Sanchez, D. Wang, I. Jungreis, C. Tomkins-Tinch, A. Kislyuk, M. F. Lin, S. Chapman, B. Macinnis, A. Matthews, J. Bochicchio, L. E. Hensley, J. H. Kuhn, C. Nusbaum, J. S. Schieffelin, B. W. Birren, M. Forget, S. T. Nichol, G. F. Palacios, D. Ndiaye, C. Happi, S. M. Gevao, M. A. Vandi, B. Kargbo, E. C. Holmes, T. Bedford, A. Gnirke, U. Ströher, A. Rambaut, R. F. Garry, P. C. Sabeti, Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone, *Cell* **161**, 1516–1526 (2015).

273. G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey, A. Rambaut, Virus genomes reveal factors that spread and sustained the Ebola epidemic, *Nature* **544**, 309–315 (2017).
274. W. E. Diehl, A. E. Lin, N. D. Grubaugh, L. M. Carvalho, K. Kim, P. P. Kyawe, S. M. McCauley, E. Donnard, A. Kucukural, P. McDonel, S. F. Schaffner, M. Garber, A. Rambaut, K. G. Andersen, P. C. Sabeti, J. Luban, Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic, *Cell* **167**, 1088-1098.e6 (2016).
275. R. A. Urbanowicz, C. P. McClure, A. Sakuntabhai, A. A. Sall, G. Kobinger, M. A. Müller, E. C. Holmes, F. A. Rey, E. Simon-Loriere, J. K. Ball, Human Adaptation of Ebola Virus during the West African Outbreak, *Cell* **167**, 1079-1087.e5 (2016).
276. J. Thézé, T. Li, L. du Plessis, J. Bouquet, M. U. G. Kraemer, S. Somasekar, G. Yu, M. de Cesare, A. Balmaseda, G. Kuan, E. Harris, C. hsi Wu, M. A. Ansari, R. Bowden, N. R. Faria, S. Yagi, S. Messenger, T. Brooks, M. Stone, E. M. Bloch, M. Busch, J. E. Muñoz-Medina, C. R. González-Bonilla, S. Wolinsky, S. López, C. F. Arias, D. Bonsall, C. Y. Chiu, O. G. Pybus, Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico, *Cell Host Microbe* **23**, 855-864.e7 (2018).
277. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK, *Science (80-.).* **371**, 708–712 (2021).
278. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America, *Science (80-.).* **370**, 564–570 (2020).
279. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M. L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, J. Hadfield, E. B. Hodcroft, J.

Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington state, *Science (80-.).* **370**, 571–575 (2020).
280. J. L. Geoghegan, X. Ren, M. Storey, J. Hadfield, L. Jelley, S. Jefferies, J. Sherwood, S. Paine, S. Huang, J. Douglas, F. K. Mendes, A. Sporle, M. G. Baker, D. R. Murdoch, N. French, C. R. Simpson, D. Welch, A. J. Drummond, E. C. Holmes, S. Duchêne, J. de Ligt, Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand, *Nat. Commun.* **11**, 1–7 (2020).
281. D. S. Candido, I. M. Claro, J. G. de Jesus, W. M. Souza, F. R. R. Moreira, S. Dellicour, T. A. Mellan, L. du Plessis, R. H. M. Pereira, F. C. S. Sales, E. R. Manuli, J. Thézé, L. Almeida, M. T. Menezes, C. M. Voloch, M. J. Fumagalli, T. M. Coletti, C. A. M. da Silva, M. S. Ramundo, M. R. Amorim, H. H. Hoeltgebaum, S. Mishra, M. S. Gill, L. M. Carvalho, L. F. Buss, C. A. Prete, J. Ashworth, H. I. Nakaya, P. S. Peixoto, O. J. Brady, S. M. Nicholls, A. Tanuri, Á. D. Rossi, C. K. V. Braga, A. L. Gerber, A. P. C. de Guimarães, N. Gaburo, C. S. Alencar, A. C. S. Ferreira, C. X. Lima, J. E. Levi, C. Granato, G. M. Ferreira, R. S. Francisco, F. Granja, M. T. Garcia, M. L. Moretti, M. W. Perroud, T. M. P. P. Castiñeiras, C. S. Lazari, S. C. Hill, A. A. de Souza Santos, C. L. Simeoni, J. Forato, A. C. Sposito, A. Z. Schreiber, M. N. N. Santos, C. Z. de Sá, R. P. Souza, L. C. Resende-Moreira, M. M. Teixeira, J. Hubner, P. A. F. Leme, R. G. Moreira, M. L. Nogueira, N. M. Ferguson, S. F. Costa, J. L. Proenca-Modena, A. T. R. Vasconcelos, S. Bhatt, P. Lemey, C. H. Wu, A. Rambaut, N. J. Loman, R. S. Aguiar, O. G. Pybus, E. C. Sabino, N. R. Faria, Evolution and epidemic spread of SARS-CoV-2 in Brazil, *Science (80-.).* **369**, 1255–1260 (2020).
282. M. U. G. Kraemer, V. Hill, C. Ruis, S. Dellicour, S. Bajaj, J. T. McCrone, G. Baele, K. V. Parag, A. L. Battle, B. Gutierrez, B. Jackson, R. Colquhoun, Á. O'Toole, B. Klein, A. Vespignani, E. Volz, N. R. Faria, D. M. Aanensen, N. J. Loman, L. Du Plessis, S. Cauchemez, A. Rambaut, S. V. Scarpino, O. G. Pybus, Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence, *Science (80-.).* **373**, 889–895 (2021).
283. M. K. Njenga, L. Nderitu, J. P. Ledermann, A. Ndirangu, C. H. Logue, C. H. L. Kelly, R. Sang, K. Sergon, R. Breiman, A. M. Powers, Tracking epidemic Chikungunya virus into the Indian Ocean from East Africa, *J. Gen. Virol.* **89**, 2754–2760 (2008).
284. J. Xavier, M. Giovanetti, V. Fonseca, J. Thézé, T. Gräf, A. Fabri, J. G. de Jesus, M. C. L. de Mendonça, C. D. dos Santos Rodrigues, M. A. Mares-Guia, C. C. dos Santos, S. F. de Oliveira Tosta, D. Candido, R. M. R. Nogueira, A. L. de Abreu, W. K. Oliveira, C. F. Campelo de Albuquerque, A. Chieppe, T. de Oliveira, P. Brasil, G. Calvet, P. C. Sequeira, N. R. Faria, A. M. B. de Filippis, L. C. J. Alcantara, Circulation of chikungunya virus East/Central/ South African lineage in Rio de Janeiro, Brazil, *PLoS One* **14**, 1–14 (2019).
285. D. Camacho, J. Reyes, A. Negredo, L. Hernández, M. Sánchez-Seco, G. Comach, Asian genotype of Chikungunya virus circulating in Venezuela during 2014, *Acta Trop.* **174**, 88–90 (2017).

286. A. Moreira-Soto, I. de O. Carneiro, C. Fischer, M. Feldmann, B. M. Kümmerer, N. S. Silva, U. G. Santos, B. F. de C. D. Souza, F. de A. Liborio, M. M. Valença-Montenegro, P. de O. Laroque, F. R. da Fontoura, A. V. D. Oliveira, C. Drosten, X. de Lamballerie, C. R. Franke, J. F. Drexler, Limited Evidence for Infection of Urban and Peri-urban Nonhuman Primates with Zika and Chikungunya Viruses in Brazil, *mSphere* **3** (2018), doi:10.1128/msphere.00523-17.
287. J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system, *Nat. Rev. Genet.* **19**, 9–20 (2018).
288. R. R. Miller, V. Montoya, J. L. Gardy, D. M. Patrick, P. Tang, Metagenomics for pathogen detection in public health, *Genome Med.* **5**, 81 (2013).
289. R. J. Rockett, A. Arnott, C. Lam, R. Sadsad, V. Timms, K. A. Gray, J. S. Eden, S. Chang, M. Gall, J. Draper, E. M. Sim, N. L. Bachmann, I. Carter, K. Basile, R. Byun, M. V. O'Sullivan, S. C. A. Chen, S. Maddocks, T. C. Sorrell, D. E. Dwyer, E. C. Holmes, J. Kok, M. Prokopenko, V. Sintchenko, Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling, *Nat. Med.* **26**, 1398–1404 (2020).
290. M. Cyrus Maher, I. Bartha, S. Weaver, J. Di Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, D. L. Robertson, G. Snell, D. Corti, H. W. Virgin, S. L. Kosakovsky Pond, A. Telenti, Predicting the mutational drivers of future SARS-CoV-2 variants of concern, *Sci. Transl. Med.* **First rele** (2022), doi:https://doi.org/10.1126/scitranslmed.abk3445.
291. A. Kalkauskas, U. Perron, Y. Sun, N. Goldman, G. Baele, S. Guindon, N. De Maio, Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk, *PLoS Comput. Biol.* **17**, 1–27 (2021).
292. K. V. Parag, L. Du Plessis, O. G. Pybus, Jointly inferring the dynamics of population size and sampling intensity from molecular sequences, *Mol. Biol. Evol.* **37**, 2414–2429 (2020).
293. P. Lemey, S. L. Hong, V. Hill, G. Baele, C. Poletto, V. Colizza, Á. O'Toole, J. T. McCrone, K. G. Andersen, M. Worobey, M. I. Nelson, A. Rambaut, M. A. Suchard, Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2, *Nat. Commun.* **11**, 1–14 (2020).
294. L. A. Featherstone, F. Di Giallonardo, E. C. Holmes, T. G. Vaughan, S. Duchêne, Infectious disease phylodynamics with occurrence data, *Methods Ecol. Evol.* **12**, 1498–1507 (2021).
295. A. E. Zarebski, L. du Plessis, K. V Parag, A computationally tractable birth-death model that 1 combines phylogenetic and epidemiological data, *bioRxiv* , 2020.10.21.349068 (2020).
296. A. M. Wensing, V. Calvez, H. F. Günthard, V. A. Johnson, R. Paredes, D. Pillay, R. W. Shafer, D. D. Richman, 2017 update of drug resistance mutatons in HIV-1, *Top. Antivir. Med.* **24**, 132–141 (2017).
297. T. Bhattacharya, M. Daniels, D. Heckerman, B. Foley, N. Frahm, C. Kadie, J. Carlson, K. Yusim, B. McMahon, B. Gaschen, S. Mallal, J. I. Mullins, D. C. Nickle, J. Herbeck, C. Rousseau, G. H. Learn, T. Miura, C. Brander, B. Walker, B. Korber, Founder effects in the assessment of HIV polymorphisms and HLA allele associations, *Science (80-.).* **315**, 1583–1586 (2007).

298. L. V. Wain, E. Bailes, F. Bibollet-Ruche, J. M. Decker, B. F. Keele, F. Van Heuverswyn, Y. Li, J. Takehisa, E. M. Ngole, G. M. Shaw, M. Peeters, B. H. Hahn, P. M. Sharp, Adaptation of HIV-1 to Its Human Host, *Mol. Biol. Evol.* **24**, 1853–1860 (2007).
299. F. Bibollet-Ruche, A. Heigele, B. F. Keele, J. L. Easlick, J. M. Decker, J. Takehisa, G. Learn, P. M. Sharp, B. H. Hahn, F. Kirchhoff, Efficient SIVcpz replication in human lymphoid tissue requires viral matrix protein adaptation, *J. Clin. Invest.* **122**, 1644–1652 (2012).
300. C. Troupin, L. Dacheux, M. Tanguy, C. Sabeta, H. Blanc, C. Bouchier, M. Vignuzzi, S. Duchene, E. C. Holmes, H. Bourhy, Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts, *PLoS Pathog.* **12**, 1–20 (2016).
301. C. Lebarbenchon, D. E. Stallknecht, Host shifts and molecular evolution of H7 avian influenza virus hemagglutinin, *Virol. J.* **8**, 328 (2011).
302. D. J. Hulse, R. G. Webster, R. J. Russell, D. R. Perez, Molecular Determinants within the Surface Proteins Involved in the Pathogenicity of H5N1 Influenza Viruses in Chickens, *J. Virol.* **78**, 9954–9964 (2004).
303. B. Gutierrez, M. Escalera-Zamudio, O. G. Pybus, Parallel molecular evolution and adaptation in viruses, *Curr. Opin. Virol.* **34**, 90–96 (2019).
304. E. L. Wise, S. Márquez, J. Mellors, V. Paz, B. Atkinson, B. Gutierrez, S. Zapata, J. Coloma, O. G. Pybus, S. K. Jackson, G. Trueba, G. Fejer, C. H. Logue, S. T. Pullan, Oropouche virus cases identified in Ecuador using an optimised qrt-pcr informed by metagenomic sequencing, *PLoS Negl. Trop. Dis.* **14**, 1–15 (2020).
305. W. Silva-Caso, M. A. Aguilar-Luis, C. Palomares-Reyes, F. Mazulis, C. Weilg, L. J. del Valle, J. Espejo-Evaristo, F. Soto-Febres, J. Martins-Luna, J. del Valle-Mendoza, First outbreak of Oropouche Fever reported in a non-endemic western region of the Peruvian Amazon: Molecular diagnosis and clinical characteristics, *Int. J. Infect. Dis.* **83**, 139–144 (2019).
306. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, *Nat. Microbiol.* **5**, 1403–1407 (2020).
307. B. Jackson, M. F. Boni, M. J. Bull, A. Colleran, R. M. Colquhoun, A. C. Darby, S. Haldenby, V. Hill, A. Lucaci, J. T. McCrone, S. M. Nicholls, Á. O'Toole, N. Pacchiarini, R. Poplawski, E. Scher, F. Todd, H. J. Webster, M. Whitehead, C. Wierzbicki, N. J. Loman, T. R. Connor, D. L. Robertson, O. G. Pybus, A. Rambaut, Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic, *Cell* , 1–10 (2021).

Published section
*Section 1.3.2.5 (*Using parallel evolution as a tool to study virus evolution*) has been published and is covered in full detail in the following review:

Gutierrez B, Escalera-Zamudio M, Pybus OG (2019) Parallel molecular evolution and adaptation in viruses. Current Opinion in Virology 34:90-96. (303)

2 The evolutionary dynamics of Oropouche Virus in South America

Highlight: Shining a light on *unsampled* lineages by identifying genomic reassortment

The following chapter presents a comprehensive evolutionary analysis of an Orthobunyavirus with a tripartite genome which has circulated in South America over several decades. The low number of sequences suggests that the virus is probably unsampled, but an additional line of evidence reveals a phenomenon which pinpoints to these unsampled lineages. As we reconstruct the widely differing ages of the three viral segments, we propose that these can only be reconciled if we account for the unsampled, unobserved lineages which "fill the gaps" on the full phylogenetic reconstruction of the viral history. As such, we infer the necessary existence of these unsampled lineages to make sense of specific patterns observed from our phylogenetic analyses.

Abstract

The Amazon basin is home to numerous arthropod-borne viral pathogens that cause febrile disease in humans. Among these, *Oropouche orthobunyavirus* (OROV) is a relatively understudied member of the genus *Orthobunyavirus*, family *Peribunyaviridae*, that causes periodic outbreaks in human populations in Brazil and other South American countries. Although several studies have described the genetic diversity of the virus, the evolutionary processes that shape the OROV genome remain poorly understood. Here we present a comprehensive study of the genomic dynamics of

OROV that encompasses phylogenetic analysis, evolutionary rate estimates, inference of natural selective pressures, recombination and reassortment, and structural analysis of OROV variants. Our study includes all available published sequences, as well as a set of new OROV genomes sequences obtained from patients in Ecuador, representing the first set of genomes from this country. Our results show differing evolutionary processes on the three segments that comprise the viral genome. We infer differing times of the most recent common ancestors (TMRCAs) of the genome segments and propose this can be explained by cryptic reassortment. We also present the discovery of previously unobserved putative N-linked glycosylation sites, and codons that evolve under positive selection on the viral surface proteins, and discuss the potential role of these features in the evolution of OROV through a combined phylogenetic and structural approach.

Importance

The emergence and re-emergence of pathogens such as Zika virus (ZIKV), Chikungunya virus (CHIKV) and yellow fever virus (YFV) have drawn attention towards other co-circulating arboviruses in South America. Oropouche virus (OROV) is a poorly-studied pathogen responsible for over a dozen outbreaks since the early 1960s, and represents a public health burden to countries such as Brazil, Panama and Peru. OROV is likely underreported as its symptomatology can be easily confounded with other febrile illnesses (e.g. dengue fever and leptospirosis), and point-of-care testing for the virus is still uncommon. With limited data, there is a need to optimise the information currently available. Analysis of OROV genomes can help us understand how the virus circulates in nature and can reveal the evolutionary forces that shape the

genetic diversity of the virus, which has implications for molecular diagnostics and the design of potential vaccines.

Introduction

The Bunyavirales is a highly diverse order of viruses that include multiple emerging human pathogens. Within this order, the *Orthobunyavirus* genus (family: *Peribunyaviridae*) includes many arthropod-borne virus species that have been associated with disease in humans (1). The severity of disease varies and ranges from self-limiting febrile illness to encephalitis and haemorrhagic fever (2). One prominent member of this genus is Oropouche virus (OROV), a common causal agent of febrile disease in the Amazon basin (3) and a potential candidate for future emerging epidemics in the region and elsewhere (4).

Since its first description in 1961 in Trinidad and Tobago (5), OROV has caused several outbreaks and sporadic infections in the Brazilian Amazon, particularly in the states of Pará, Amapá, Rondônia, Maranhão, Acre, Amazonas, Minas Gerais, Mato Grosso, and Tocantins (6-9), and evidence suggests the circulation of OROV in other Brazilian states (10). Since the late 1980s, additional cases and outbreaks of OROV have been reported in Panama, Peru (1113) and, more recently, Ecuador (14). The virus has been isolated from sloths (*Bradypus trydactilus*) (3) and marmosets (*Callithrix sp.*) (15) as well as from invertebrates, such as *Culex p. quinquefasciatus* (9) and *Ochlerotatus serratus* mosquitoes and *Culicoides paraensis* biting midges (3). It has been proposed that OROV has two distinct transmission cycles (16). In the sylvatic transmission cycle, wild mammals such as sloths (*B. trydactilus*) and primates (*Allouatta caraya*, *Callithrix*

penicillata) serve as hosts for the virus, which is transmitted by vector species that are common in rural areas (*Aedes serratus* and *Coquillettidia venezuelensis* have been proposed as vectors in the sylvatic cycle). This sylvatic cycle may possibly include other vertebrate hosts, including wild birds and rodents (such as *Proechimys sp.*). In contrast, the urban transmission cycle involves human-to-human transmission mediated by the biting midge *C. paraensis* (16) and can be facilitated by anthropogenic disturbance of forest areas.

OROV produces an acute febrile illness in humans, characterised by unspecific symptoms such as fever, chills, headaches, myalgia and joint pain (17) and is therefore difficult to diagnose and differentiate from other co-occurring infectious diseases such as leptospirosis, dengue fever, Venezuelan Equine Encephalitis (VEE), malaria, and rickettsial and *Coxiella* infections (18). Specific to the Amazon region, OROV infections are difficult to distinguish from undifferentiated febrile illness associated to group C orthobunyaviruses (e.g. Itaqui and Caraparú viruses) or phleboviruses from the Candiru complex.

The genome of OROV is typical of the orthobunyaviruses and comprises three negative-sense RNA segments of different sizes: a large segment (L) that encodes a RNA dependent RNA polymerase, a medium (M) segment that encodes a membrane polyprotein comprised of three main components (the Gn, NSm and Gc proteins), and a small (S) segment that encodes the structural nucleocapsid protein N and a second non-structural protein, Ns, in an overlapping reading frame (2). The Gn and Gc proteins determine the architecture of the viral particle; Cryo-EM studies show that the

membrane proteins of Bunyamwera virus (BUNV) form a tripodal structure (19) and the crystal structures of the N-terminus of Gc suggest that this trimeric assembly may be conserved across various orthobunyaviruses (20). The segmented nature of the OROV genome enhances the likelihood of genome reassortment events, a phenomenon that has been observed more generally in the bunyaviruses (2) and has been associated, in some cases, with a dramatic increase in disease severity, such as with Ngari virus, a reassortant between BUNV and Batai virus (BATV) (21). Furthermore, reassortment appears to play an important role in the evolution and emergence of new viruses within OROV and related species, as has been described for the reassortant Jatobal virus (JATV) (22) and Perdoes virus (7) in Brazil, Iquitos virus (IQTV) in Peru (23) and Madre de Dios virus (MDDV) in Venezuela (24).

Despite a heightened interest in OROV, reflected by an increasing number of publications covering the virus in recent years (16), many aspects of OROV history, biology and ecology remain poorly understood. Although the evolution of the virus has been previously studied using standard phylogenetic methods (3, 7, 8, 25, 26), a comprehensive analysis of the molecular evolution of OROV that incorporates the complete genetic diversity of the virus (represented by whole genomes) and a comparison of the evolutionary histories of its genome segments has not yet been undertaken. Here we present the results of an in-depth analysis of the OROV genomes from the Americas that includes the genomes of 6 new OROV isolates sampled in Ecuador (27) and explores the differences between the evolutionary histories of the three viral segments through comparative phylogenetics. We evaluate the temporal signal of

each RNA segment, estimate their dates of origin and rates of evolution, and test for recombination and reassortment. Furthermore, we combine phylogenetic tests for natural selection on the viral genome with structural mapping and N-linked glycosylation sequon analysis to determine the factors that drive the evolution of the viral genome and explain the observed differences between segments.

Methods

Sampling and sequencing of OROV isolates from Ecuador

Six strains of OROV were isolated independently via one passage in Vero cells, from febrile patient sera sampled in 2016 in the Esmeraldas region of Ecuador (27). RNA was extracted and six complete genome sequences were generated using a metagenomic approach as described previously (14). Briefly, cDNA was prepared using a Sequence Independent Single Primer Amplification (SISPA) approach (28), cDNA sequencing libraries were prepared using the Nextera XT Kit (Illumina), and sequencing was performed on a MiSeq (Illumina) using 150 nucleotide paired-end runs. Reads were mapped to reference sequences using BWA MEM (29). The final consensus sequences were generated using Quasibam (30). We visually inspected each data set for the occurrence of single nucleotide variants (SNVs) within each patient; limited variation per site was found for all isolates (detailed in Supplementary File 1). The study was approved by the bioethics committee of Universidad San Francisco de Quito, and all patients provided written consent indicating that they agreed for their samples to be tested for additional pathogens.

Data sets

In addition to the Ecuadorian sequences, we collected all available sequences for the three OROV genome segments from GenBank at NCBI. We then excluded all sequences that were shorter than a threshold segment length (specifically, <6500 nucleotides long for L, <4200 nucleotides long for M and <693 nucleotides long for S; these lengths correspond to the complete coding regions of the genes contained within each segment). These GenBank sequences represent the genetic diversity of OROV from outbreaks in Trinidad & Tobago, Panama, Peru and Brazil, and cover the complete sampled history of the virus from the mid 1950s to the late 2000s, with the latest outbreak in Brazil reported in 2009 (7). The data sets were compiled into individual alignments for each of the three viral genome segments (denoted L for large, M for medium and S for small). A complete list of the GenBank accession numbers of all sequences used is presented in Table S1.

Additional data for the OROV sequences included the sampling date (with varying precision of year, month or exact day, depending on the sequence), the identity of the host species (most samples were obtained from human patients, while a few were obtained from potential arthropod vectors and mammalian reservoir species), and the location of each sample (province of origin was used for the samples from Brazil and Peru, while only the country of origin was available for the Panama sequences; samples from Ecuador and Trinidad & Tobago were obtained from a single location). No exact

sampling date is available for the Ecuadorian sequences, but sampling was conducted between the months of April and May of 2016 (Sully Márquez, Pers. Comm.).

Before sequence alignment, the untranscribed terminal repeats (UTRs) of each segment were removed and the open reading frame of each segment was identified (only the largest ORF of the S segment, corresponding to the N gene, was analysed). Sequences were then aligned using the MUSCLE algorithm, as implemented in Geneious 9.0.5 (31), and the alignments were manually checked and edited to ensure they were codon-aligned. The L alignment included 59 published sequences plus the six new sequences from Ecuador. The M segment alignment included 58 published sequences and the six Ecuadorian sequences. Finally, the S segment alignment included 143 published sequences plus the six Ecuadorian sequences.

Molecular phylogenetics and molecular clock analyses

All three alignments were scanned for recombinant sequences using the suite of methods implemented in RDP4 (RDP, BOOTSCAN, MAXCHI, CHIMAERA, 3SEQ, GENECONV, LARD and SISCAN) (32). Sequences identified as recombinant (i.e. evidence for recombination was found by at least four methods) were removed from all further analyses (only two M segment sequences were excluded). For each alignment, a molecular phylogeny was estimated using the maximum likelihood (ML) approach implemented in RAxML 8.0 (33). ML trees were estimated using a General Time Reversible (GTR) substitution model with a Gamma-distribution model of among-site heterogeneity. Phylogenetic node support was assessed using a non-parametric bootstrap approach with 100 replicates. No outgroup sequences were used and all trees were

midpoint rooted. A visual comparison of the tree topologies (after collapsing some nodes to facilitate visualisation) and bootstrap scores for each segment was used to evaluate possible OROV reassortment events during its evolutionary history. Well supported topological incongruencies between trees were interpreted as a potential reassortment event between the corresponding genome segments.

To explore the rates of OROV molecular evolution and to evaluate the temporal signal in each OROV alignment, we correlated tip-to-root genetic distances in the ML tree against the sampling dates of the corresponding sequences, using the approach implemented in TempEst 1.5.1 (34). The regression plots were inspected visually to identify notable outliers and the linear correlation coefficient of the regression was used as a measure of the degree to which the sequences evolve in a clock-like manner.

We estimated the evolutionary history of OROV outbreaks in South America by constructing time-calibrated Maximum Clade Credibility (MCC) trees for each segment. Each MCC tree summarises a posterior distribution of phylogenies that were sampled using the Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in BEAST 1.10.1 (35). Trees were sampled under a HKY substitution model with codon-position partitioning and a Gamma distribution model of among-site heterogeneity (this substitution model was chosen following the results of (36); this model differs from the aforementioned and more parametrised GTR model used for the ML analyses to avoid the matrix exponentiation approximations required for the usage of this model in a Bayesian framework), a Bayesian Skyline tree prior (37), and a relaxed uncorrelated molecular clock (UCLD) (38) model. For all molecular clock analyses from this point

onward, the outlier sequences identified from the tip-to-root regression plots (Fig S1) were removed from their corresponding data sets.

Our ML and TempEst analyses (above) revealed two monophyletic lineages in the M segment phylogeny, one with weak temporal signal (correlation coefficient=0.66) and one with stronger signal (correlation coefficient=0.94; see *Results* and Fig S1). These lineages were separated by long internal branches that might have unexpected effects on the estimation of evolutionary rates and the age of the root of the tree (39). Therefore, the molecular clock analysis of the M segment was performed as a two-step process. First, we estimated a time-calibrated tree separately for the two M lineages and noted the estimated time of the most recent common ancestor (TMRCA) of the larger Lineage 1 (the very small sample size of Lineage 2 means it is unlikely to generate a robust evolutionary timescale; see results below and Fig S5). We then performed a second analysis of the whole group M phylogeny, in which this TMCRA estimate was used as an additional calibration point (i.e. as an informative prior on the date of the common ancestor of Lineage 1). Similarly, mean evolutionary rates were co-estimated with phylogeny for each of the three segments using the UCLD clock model and subsequently compared. We further repeated this analysis separately for M segment Lineages 1 and 2. Additional analyses using a less realistic strict molecular clock were also performed for model comparison purposes. All XML files are available via GitHub at https://github.com/BernardoGG/OROV_America.

Selection analyses

The ORF alignments of the S and L segments, as well as those of the individual proteins of the M segment ORF (Gn, NSm and Gc) were tested for evidence of natural selection using the various methods implemented in HyPhy (40), implemented by the Datamonkey 2.0 server (41). Statistical tests were used to evaluate evidence for both pervasive and episodic selection in the OROV genome, and to evaluate selection across whole genes as well as at specific codons. Analyses of pervasive selection were performed using three methods: (i) the Branch-site Unrestricted Statistical Test for Episodic Diversification (BUSTED) method, which tests for gene-wise selection and estimates a mean *dN*/*dS* ratio for each gene (42), (ii) the Single Likelihood Ancestor Counting (SLAC) method, and (iii) the Fixed Effects Likelihood (FEL) method; the latter two approaches were used to infer specific codons under adaptive selection (43). Episodic selection was evaluated using two methods: (i) the Mixed Effects Model of Evolution (MEME), which identifies individual codons under positive selection (44), and (ii) a branch-site model implemented in aBSREL (45), which was used to search for phylogenetic branches under selection. A final analysis was performed using the Fast Unconstrained Bayesian Approximation (FUBAR) (46) method to evaluate the difference between nonsynonymous (b) and synonymous rates (a) per site for each viral segment. All codon positions were numbered relative to the OROV reference strain BeAn19991.

Structural analysis of OROV proteins and sites under selection

The availability of high-resolution protein structures for the OROV proteins is limited. We therefore focused efforts on the OROV glycoprotein Gc because (i) N-terminal regions for a number of orthobunyaviral Gc glycoproteins (including the head domain of OROV Gc, PDB ID 6H3X), have been structurally elucidated (20) and (ii) this protein shows an increased number of sites under positive selection compared to other OROV proteins (see Results). The second component of the OROV glycoprotein, Gn, was not analysed here because no adequate atomic resolution structure is currently available. For Gc, N-linked glycosylation sites were predicted for each virus sequence using the NetNGlyc 1.0 server (www.cbs.dtu.dk/services/NetNGlyc/), which identifies NXT/S amino acid motifs (where X is any amino acid except proline). The sequence evolution of these motifs was subsequently mapped onto the M segment phylogeny using a maximum parsimony approach. To enable structure-based mapping of sites under selection, we created a composite molecular model that encompasses the structurally elucidated N-terminal head of OROV Gc (PDB ID 6H3X) and the stem domains from the closely related Schmallenberg virus (SBV) (PDB ID 6H3S) (20) through sequence-independent structural alignments in PyMOL (47). As there are no known structures of the orthobunyaviral C-terminal domains (henceforth called the core region), this region was omitted from the structural analysis. Residues in the N-terminal head of the OROV Gc structure that contribute to the crystal packing-induced homotrimeric surface (20) were identified with the PDBePISA server (48). Sites under selection were mapped onto the composite and the trimeric models, and their residue-specific solvent accessibility was analysed using the ESPript 3.0 server (49).

We explored how the evolutionary history of OROV has been shaped by molecular adaptation of different domains of the Gc protein by estimating *dN*/*dS* ratios for each protein domain using the renaissance counting approach (50) implemented in BEAST 1.10.1 (35). These estimates were independently performed for the full set of M sequences and for the two previously described lineages, using an empirical posterior sample of phylogenies comprising >100,000 trees. The 95% highest posterior distribution (HPD) credible intervals for the normalised *dN* and *dS* values were summarised using the *coda* package in R (https://cran.r-project.org/web/packages/coda/index.html). The normalised values are the observed count of each substitution type, divided by the expected count of each substitution type under the substitution and evolutionary models and independent of the data.

Results

Molecular phylogenetics and molecular clock analyses

The estimated ML phylogenies for each OROV genome segment are shown in Figure 1. An interactive phylogenetic representation that includes additional information such as sampling location is available through Microreact (51) (URLs provided in Table S2). The phylogenies show spatially structured virus populations as evidenced by the clustering of taxa according to country of isolate collection. In all three phylogenies (Fig 1a, b, c), our new sequences from Ecuador (yellow) cluster together in a single monophyletic group (with 100% bootstrap support) that descends from OROV sequences from Peru (red), suggesting the Ecuadorian sequences represent an

unreported outbreak in 2016 that is related to cases observed in Peru in the mid-to-late 2000s. The large (L) and medium (M) segment trees (Fig 1b, c) contain long internal branches and distinct lineages. Sequences from Panama (orange) are found in two distinct clusters, while Brazilian sequences (light blue) are split into two distinct groups. One Brazilian group corresponds to the 2009 OROV outbreak in the municipality of Mazagão (Amapá state) in northern Brazil (7), whilst the other contains sequences from isolates obtained between 1960 and 2006 from seven different states (Acre, Amapá, Amazonas, Maranhão, Minas Gerais, Pará and Rondônia) (52). Most of these clusters are supported by high bootstrap scores (>95%) but there is less consistent and sometimes limited support for phylogenetic structure *within* clusters. The OROV small (S) segment phylogeny (Fig 1a) contains more sequences but has less phylogenetic structure (i.e. internal branches are shorter and several basal nodes have weak bootstrap support). Geographical clustering is also less evident; there are 3 well supported clusters of sequences from Brazil or Brazil/Panama, the Ecuadorian sequences are again found in a strongly supported cluster descended from the Peruvian isolate IQE7894, but the remaining sequences from Peru cluster elsewhere in the tree (Fig 1).

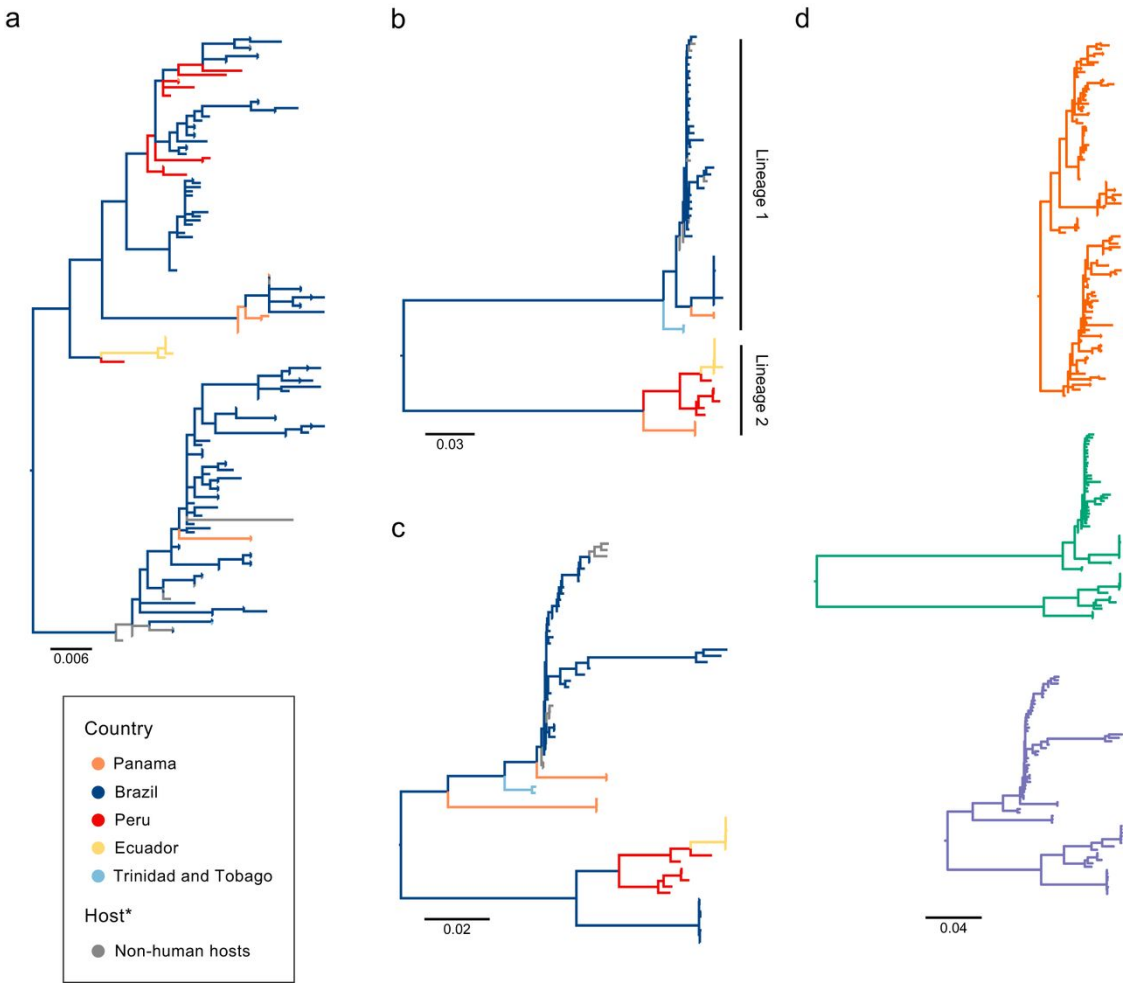


Figure 1. Maximum Likelihood trees for the three OROV genome segments.

(a) Small (S) segment phylogeny. **(b)** Medium (M) segment phylogeny. **(c)** Large (L) segment phylogeny. Branches are colour coded by the country of origin of the tips and an ancestral state reconstruction of the nodes using a parsimony approach, and is not intended as a phylogeographic analysis. Grey tips/branches indicate sequences obtained from non-human hosts and vectors (species indicated in the tip label). Node support from 100 bootstrap replicates is shown for the basal internal nodes of each tree. **(d)** Scaled comparison of the three segment phylogenies (genetic distance between vertical grid lines is 0.02) highlights the different node depths between all trees.

The trees for the different segments show considerable topological differences, suggesting the likelihood of reassortment events in the viral genome. Pairwise

comparisons between segment phylogenies (Fig S2) indicate that the L segment of the 2009 Mazagão outbreak sequences has arisen through a reassortment event. There are also widespread topological differences between the S segment and the L and M segments (Fig S2), particularly for the Brazilian sequences (which cluster in robust monophyletic clades in the L and M segment trees but as multiple clusters in the S segment phylogeny). This suggests that OROV might reassort frequently; however, the low node support for the S segment could also represent phylogenetic uncertainty rather than differing evolutionary history. Furthermore, a comparison of the three segment trees drawn on the same scale shows widely differing tree lengths (Fig 1d).

The evolutionary timescale for OROV in America, estimated using a Bayesian phylogenetic molecular clock approach (relaxed UCLD clock model), yielded highly variable dates for the most recent common ancestor (TMRCA) for each segment. The origins of the S and L segments can be traced back to the early to mid 20th century, while the M segment shows long internal branches and an older mean root age, dating back to approximately 1608 (95% HPD: 1116-1924) (Fig 2a). The estimated TMRCA of the S segment is approximately 1940 (95% HPD: 1916-1955), while that of the L segment is slightly older, around 1919 (95% HPD: 1854-1955). It is also noteworthy that the M segment splits into two distinct lineages (Fig 2a), observed both in the ML and the Bayesian phylogenies. Lineage 1 represents an eastern-central clade that includes the earliest OROV case in Trinidad and Tobago, sequences obtained from patients in Panama between the late 80s and late 90s, and the entirety of the Brazilian genetic diversity of the virus. Lineage 1 is estimated to have emerged around 1887 (95% HPD:

1804-1945). On the other hand, Lineage 2 contains sequences from western South America, including sequences from Panama (from the late 1980s), Peru and Ecuador, with an estimated origin around 1905 (95% HPD: 1841-1962). The estimated mean evolutionary rate of the M segment appears to be significantly lower than that of other segments but still within biologically feasible values (mean rate = $1.75x10^{-3}$ subst./site/year for segment S, $3.65x10^{-4}$ subst./site/year for segment M, and $1.64x10^{-3}$ subst./site/year for segment L). We also attempted to estimate evolutionary rates estimated separately for each lineage within the M segment phylogeny. A systematic comparison of different molecular clock methods was used to evaluate the reliability of the lineage-specific estimates (Fig S4). Whilst the estimated rates for Lineage 1 are similar to those for the whole M segment, the rate estimates for Lineage 2 are highly uncertain and also inconsistent among methods and models (Fig S4). Thus, the small sample size of Lineage 2 renders the estimate for that lineage unreliable and prevents us from concluding the lineages evolve at different or similar rates. The systematic comparisons in Fig S4 and 2b also show that the relative rates of the segments are broadly consistent (i.e. L is fastest, S intermediate, and M slowest; Fig 2b and Fig S4), despite differences in their estimated absolute rates. The UCLD clock remains the most reliable and realistic model choice due to the high among lineage rate variation we observe (coefficient of variation >1 for all segments) and low R^2 values in the linear regression analyses (Fig S1).

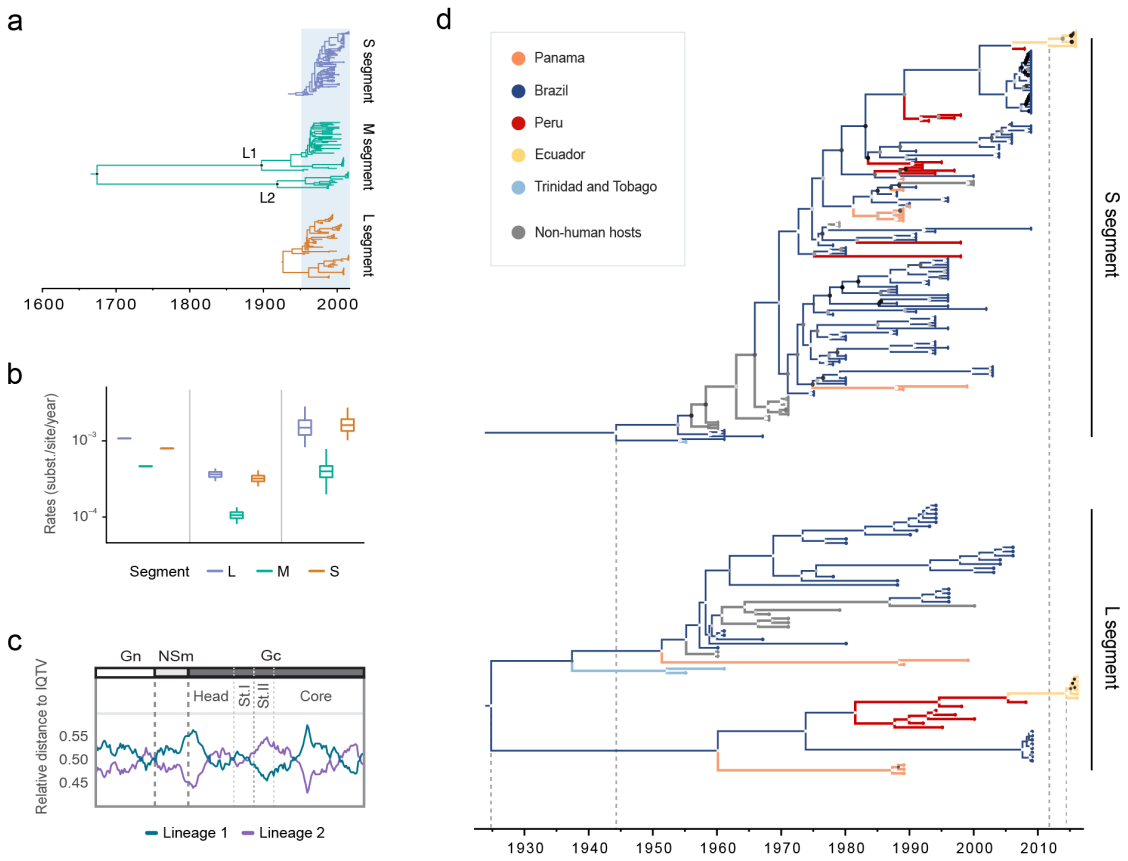


Figure 2. Time-calibrated and evolutionary rate analysis of the OROV genome.

(a) A comparison of the three segments reveals an older common ancestor for the M segment, resulting in the divergence of two lineages that appear to diversify in the early to mid 1900s (Lineage 1 and Lineage 2, L1 and L2 in the figure). **(b)** The estimated evolutionary rates of the segments. **(c)** A sliding window estimation of the relative genetic distance of each of the two M segment lineages (shown in orange and blue) to Iquitos Virus (IQTV), used as an outgroup. Points of the sequences where the closest genotype to the outgroup changes could be interpreted as recombination events with unsampled sequences. **(d)** An analysis of the younger segments, S and L, reveals emergence date estimates for the virus in the early to mid 20th century, and relatively consistent estimates for the TMRCA of the new Ecuadorian sequences in the early 2010s. The posterior probabilities (PP) of each node is shown based on a greyscale colour scheme (0.0 PP = black to 1.0 PP = white).

In order to explore whether these differences in evolutionary rates might result from rate heterogeneity between the M segment lineages, we devised a novel analysis where we calculated the mean relative genetic distance of Lineage 1 and 2 to a closely related

outgroup, Iquitos virus (IQTV), using a sliding window approach. In each window we divided the mean genetic distance between all sequences in each lineage and the outgroup by the sum of the genetic distances for both lineages, resulting in a *relative* genetic distance to IQTV (see Supplementary R Script at https://github.com/BernardoGG/OROV_America/blob/master/OROV_SimPlots.R). These relative distances are the fraction of the cumulative genetic distance of both lineages to the selected outgroup (and accordingly sum to 1). Different regions of the M segment exhibit different relative distances to the outgroup, ranging between 0.58 and 0.42 (as expected for two lineages that are roughly equidistant to the outgroup). The lines cross where one lineage becomes relatively less divergent to the outgroup and several such events occur between the NSm and Gn reading frames, within the Gn reading frame, and the core region and head domain of the Gc reading frame (Fig 2c). These differences in relative genetic distances could be the product of mechanistic phenomena, ranging from natural selection to intra-segment recombination, which would not have been detected by the RDP4 analyses because we explicitly assigned an outgroup and measure the genetic distances between lineages as their individual contributions to the cumulative distance rather than just mean genetic distances. However, it is hard to determine whether our observed values are meaningful evidence of intra-segment rate variation, and we therefore cannot conclude that such intra-segment evolutionary rate variation between the two lineages is evident.

In the S segment phylogeny, samples obtained through the 1950s and 1960s are located at the base of the tree. The tree splits into two co-circulating clusters around

1969 (95% HPD: 1964-1973). Both clusters include sequences mostly from central Brazil and Panama, and one of the clusters suggests multiple introductions to Peru as early as the mid 1970s. The 2009 Mazagão outbreak and the 2016 Ecuador sequences arise from one of these Peruvian clusters, with the latter sharing a common ancestor around late 2011 (95% HPD: 2009-2015) (Fig 2d). On the other hand, the L segment tree shows two divergent lineages that split sometime in the early to mid-twentieth century (95% HPDs: 1900-1953; 1918-1979) and more closely match the M segment lineage structure. One of these L segment lineages includes most of the Brazilian diversity (the Mazagão sequences are distinct and cluster in the other L segment lineage) and a cluster of sequences from Panama. The other includes a second cluster of Panama sequences and all of the Peruvian diversity (Fig 2d). The Ecuadorian sequences are again descended from Peruvian strains, with a TMRCA in early 2014 (95% HPD: 2011-2016). This date is similar in the other segments (the estimated Ecuador TMRCA is mid 2011 in the M segment; 95% HPD: 2008-2015). These results suggest a recent introduction of OROV to northern Ecuador with no reassortment events leading to its emergence.

Selection Analyses

We undertook a screening for sites under selection using different models implemented in the HyPhy framework (40). Evidence for both pervasive and episodic selection was found in different genes using different methods (Table 1). Pervasive selection (selection that occurs persistently along the whole phylogeny) was detected in the S segment. Individual sites under pervasive selection were identified in the M segment alignments: for the Gn protein one site was found to be evolving under positive

selection (by the SLAC and FEL methods); for the NSm protein one site was identified (by the FEL method); and the Gc protein, two sites were found to be under selection (by the FEL and FUBAR methods). Sites evolving under pervasive selection were also identified in the S and L segments (one site in each) by the FUBAR method (Table 1). Episodic adaptive selection (selection that occurs heterogeneously across the tree branches) was detected in all alignments under the mixed effects model (MEME) (44). In total, MEME identified six sites under selection in the RNA dependant RNA polymerase, seven sites under selection in the M segment polyprotein, and one site under selection in the non-structural N protein. Of these sites, only four were identified by two or more methods: codon 66 of the Gn scaffold protein, codon 86 of the NSm non-structural protein and codons 269 and 442 of the Gc spike protein (Table 1). We note that three of these four sites feature more than two alleles, and the frequencies of the most common allele range from 0.44 to 0.73 (Table S3). This episodic selection is made evident when comparing the difference between synonymous and non-synonymous rates per site on the M segment, where, despite a low proportion of sites identified as evolving under selection (Fig S5a), each lineage shows different sites with higher non-synonymous substitution rate (Fig S5b).

Structural analysis of OROV proteins and sites under selection

Viral glycoproteins are important targets of positive selection and drivers of virus adaptation due to their direct interactions with the host immune system (53). In orthobunyaviruses, glycoproteins are expressed as a polyprotein, encoded by a single open reading frame of the M segment that is expressed and post-translationally cleaved

into the Gn, NSm (non-structural) and Gc proteins respectively. The main component of the OROV particle spikes is Gc, which is composed of four main domains: a head domain, two stem domains, and a core region (20). We identified eight sites under selection in Gc overall (four of which fall within the head and stem domains) and two putative, previously unreported N-linked glycosylation sites (Fig 3a). Estimated *dN*/*dS* ratios are similar for each domain (Fig 3b), hence we find no evidence for significantly different selective pressures among domains. However, if *dN*/*dS* ratios are estimated separately for the two M segment lineages, then we obtain higher *dN*/*dS* estimates for the head and stem domains of Lineage 2, although the large confidence limits mean these differences are not significant (Fig 3b).

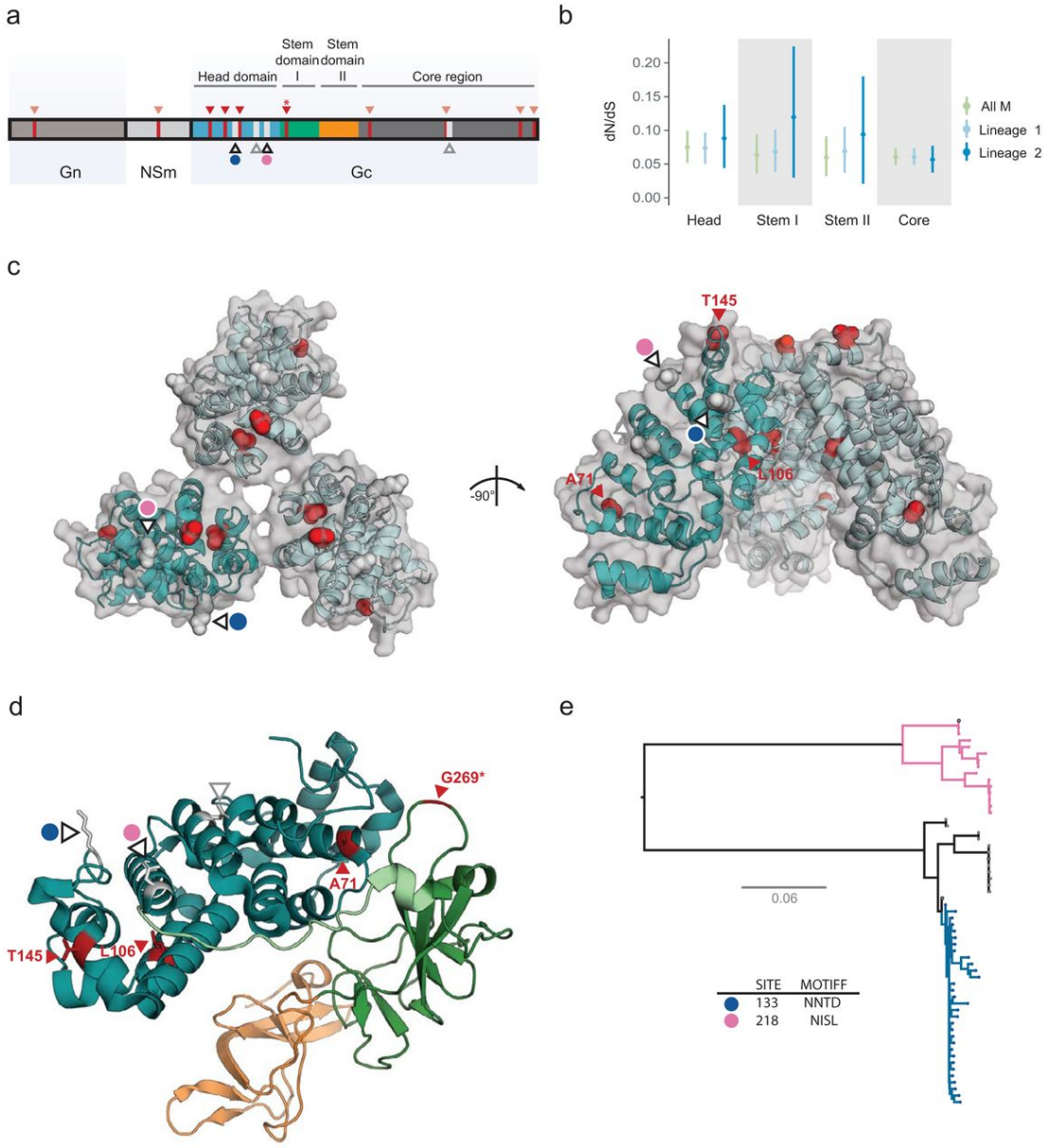


Figure 3. Structure-based mapping of sites identified under selection and previously unreported N-linked glycosylation sites onto the Gc protein.

(a) A schematic of the OROV M segment polyprotein, which includes the Gn and Gc components of the viral membrane proteins. The Gc protein is subdivided into a head domain (blue), two stalk domains (green and orange), and a C-terminal region, which encodes a putative fusion loop (dark grey). Sites identified as being under positive selection using the MEME approach are shown in red (lighter coloured arrows indicate sites falling outside of the Gc head and stem domains). Site 269 was identified as being under positive selection through multiple approaches, and marked with an asterisk. All N-linked glycosylation sites are shown in white, and the two previously undescribed sites are highlighted in blue and pink. **(b)** Estimated *dN*/*dS* ratio of each

domain in Gc, obtained for the whole M segment, and for each M segment lineage. **(c)** Crystallographically-observed trimeric spike structure of the Gc head domain, with sites under positive selection shown in red. The cartoon representation of each chain is embedded in a transparent surface representation (grey). Glycosilation sites are shown as in panel (a). **(d)** Sites under positive selection and residues corresponding to N-linked glycosylation sites mapped to a composite model generated from previously reported OROV Gc (head domain, PDB ID 6H3X) and SBV Gc (stem domains, PDB ID 6H3S) crystal structures. The OROV head domain is shown in blue, with the C-terminus loop residues shown in light green (taken from the SBV head domain) and SBV stem domains I and II are shown in green and orange, respectively. **(e)** Putative N-linked glycosylation sites identified in this study are mapped onto the phylogeny of the M segment, showing the independent presence the two sites in the two M segment lineages.

We mapped all the sites identified as being under positive selection to the available atomic structures of the OROV Gc head domain and the closely related SBV stem domains (Fig 3b, 3c). This analysis reveals that three sites under selection (corresponding to residues A71, L106 and T145) are located at alpha-helixes in the head domain, and one (residue G269) occurs at a loop connecting the stem domain I N-terminus to the head domain (Fig 3c). While A71 is buried and residue L106 is only partially accessible from within the protein structure, residues T145 and G269 are completely accessible (Fig S3), with the former located at the top of the trimeric spike structure (Fig 3b). Structure-based mapping of the residues equivalent to T145 in BUNV and LACV suggests that this residue could help to facilitate the formation of a closed trimeric interface (20), thus suggesting that substitutions at this location may affect the protein-protein interface and change the antigenic surface of the spike complex.

In addition to the sites under positive selection, we identified two previously unreported N-linked glycosylation sequons at residues 133 (motif NNTD) and 218

(motif NISL). Both of these putative glycosylation sites are located on the Gc head domain (Fig 3a) and do not cooccur together in any of the sequences, supportive of the hypothesis that they may be alternative motifs that could serve similar functional roles. Furthermore, these sites appear to be phylogenetically associated with the different M segment lineages, with motif NNTD being exclusive to Lineage 1 and motif NISL occurring exclusively in Lineage 2 (Fig 3d). Based on shared ancestry, it is likely that the motif NNTD appeared in a Brazilian common ancestor that excludes the Mazagão outbreak sequences, as this motif is not found in either the Panama or Trinidad sequences belonging to Lineage 1.

Both the putative fusion loop in the C-terminal region of the Gc protein (54) and two previously reported N-glycosylation sites were conserved amongst all analysed OROV sequences (3).

Discussion

The evolutionary genomics of OROV in the Americas appears to be driven by a complex combination of factors that are associated with viral life cycle and the nature of the OROV genome. Phylogenetic analysis of the three OROV genome segments reveals different evolutionary histories, with a stronger topological concordance between the M and L segments, than between those two segments and the S segment. Many of these incongruencies are likely to represent the different evolutionary histories of each segment, arising from reassortment. The occurrence of reassortment events appears to be relatively common for OROV in the Americas and may be currently underestimated because (i) there are likely to be many undiscovered OROV lineages

and (ii) the virus likely circulates widely in one or more as-yet uncharacterised reservoir populations in the Amazonian forest. Although we found similarities between the M and L segment phylogenies, a previous study of the Bunyamwera group has suggested linkage between the L and S segments (55), and different reassortment patterns have been reported for other bunyaviruses such as La Crosse virus (LACV) and snowshoe hare virus (SSHV) (56). The reassortment of OROV is in line with the observation of naturally occurring reassortant viruses (23, 24), a phenomenon that has been linked to the coinfection of hosts or arthropod vectors. The latter provide a good opportunity for reassortment to occur, as an infected vector can feed on a second vertebrate host infected with a distinct viral strain within a 2-3 day time window and become co-infected (2). On a broader scale, reassortment events could play a crucial role in the evolution of bunyaviruses in general, and it has been suggested that an increasing number of the recently described species are reassortants (57).

Molecular clock phylogenetic analyses can be used to infer the emergence times of pathogens and the timescales of outbreaks (58). Some of these tools are particularly applicable to heterochronous data sets, in which sequences from rapidly evolving populations have been sampled longitudinally through time. For molecular clock analyses to be reliable, samples collected at different time points must accumulate sufficient genetic differences to be distinct from one another (59) (i.e. contain sufficient 'temporal signal'). A combination of factors may affect the accuracy of estimated TMRCAs. Our analyses of the S and L segments suggest that OROV first emerged in the early to mid 1900s, a more recent estimate than previously reported (late 1700s,

inferred from S segment sequences) (26). This difference may reflect the fact that our analysis incorporates more sequences, or it may reflect differences in the evolutionary models or prior distributions used in Bayesian phylogenetic inference. It should be also noted that the aforementioned earlier study (26) did not report a test for temporal signal in the data set; absence of this signal could affect the precision of estimates.

A key factor of OROV evolutionary dynamics is the distinction between two well supported lineages for the M segment. We estimate that these two lineages split more than two centuries before the estimated TMRCAs of the other segments. The split between Lineages 1 and 2 is also reflected in the L segment phylogeny, where two well supported lineages that mostly coincide with the M phylogeny are observed, with the exception of the 2009 Mazagão outbreak reassortant group (see Fig 2 and S2).

We sought to explain the substantially older TMRCA of the M segment compared to the other two segments, despite the fact that all segments share overlapping sets of strains. The temporal signal we observed (Fig S1) suggests that the roots of the M and L segment phylogenies represent different ancestral viruses, and that the discrepant TMRCAs are not the result of estimation error. How can this conclusion be reconciled with the topological similarities between the M and L segment phylogenies? We suggest that these results could be explained by hypothesising that the ancestor of one of the M segment lineages underwent a past reassortment event with an unsampled, divergent OROV lineage (Fig 4). At present it is not possible to determine which of the two lineages actually underwent reassortment (for visualisation in Figure 4, Lineage 2 was chosen randomly). Under this hypothesis, the TMRCAs of segments M and L can differ

substantially yet their phylogenies remain largely congruent. The L segment has a proximal origin which is closer in time to the TMRCAs of both M segment lineages, which share a more ancient origin among them. Our hypothesis posits that the sequences in our data sets are part of a larger diversity of OROV in South America which is currently unsampled. Given that reassortment is common in Orthobunyaviruses, a portion of the L sequences that correspond to one of the M segment lineages represent an instance were part of this broader unsampled viral diversity reassorted into the L segment evolutionary history, therefore giving the impression of differing origin times between segments when in fact these represent different ancestors. By extension, unsampled and more divergent L segment genomes exist which, if included in the phylogenetic tree, would also descend from the more ancient inferred M segment most recent common ancestor (Fig 4).

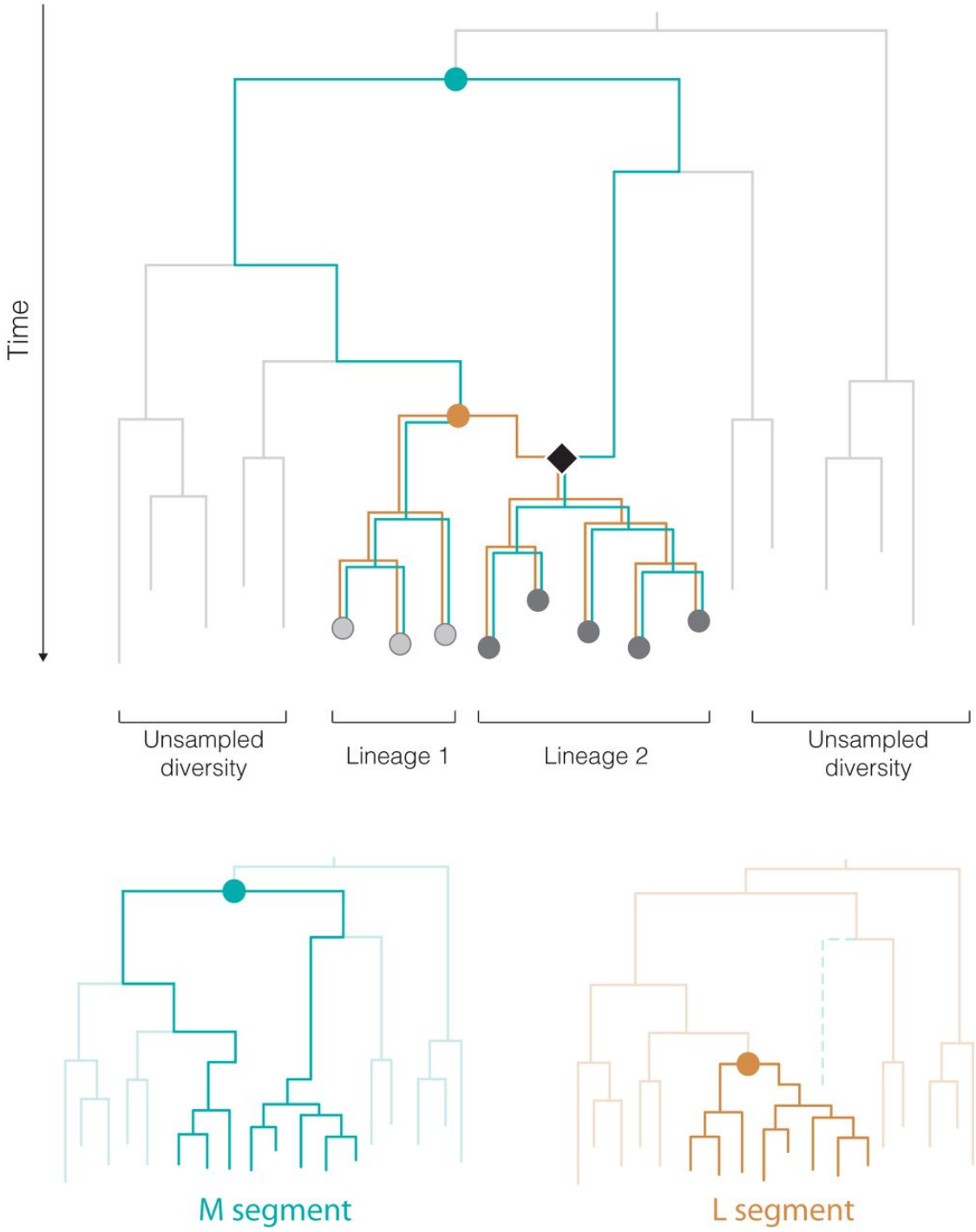


Figure 4. A model of OROV genome evolution and reassortment that can resolve the inconsistent TMRCA estimates of the L and M segments.

The upper panel depicts a simplified ancestral graph that describes the joint ancestry of the L (orange) and M (teal) segments. Sampled sequences are represented by dark and light grey circles at the tree tips. These are grouped into Lineages 1 and 2, which

correspond to the two main lineages in the M and L phylogenies (see Figure 2). The light grey branches represent hypothesised, highly diverse, unsampled OROV lineages. The graph contains one reassortment event (black diamond), resulting in the ancestor of Lineage 2 in the M segment being descended from a divergent, unsampled lineage (alternatively, the reassortment might have occurred on the branch ancestral to lineage 1). As a result, the TMRCA of the L segment phylogeny (lower right panel; orange circle) exists substantially earlier of the M segment (lower left panel; teal circle). For simplicity, other more recent reassortment events between Lineages 1 and 2 are not shown (i.e. the 2009 Mazagão outbreak clade).

The lower evolutionary rate of the M segment is also unexpected, given the tendency of Bunyaviral surface proteins to evolve faster than the nucleocapsid and non-structural proteins (60, 61). These differences are hard to reconcile with the available data, but the reliability of the temporal signal should be considered. A systematic comparison of different molecular clock approaches indicates a Lineage 2-specific rate cannot be reliably estimated. We conclude that most of the temporal signal in the M segment comes from Lineage 1 sequences, due to the small sample size and high uncertainty of Lineage 2 rate estimates. More OROV genome sequences (especially for Lineage 2) are clearly needed to provide a robust understanding of the whether the molecular evolution of the two lineages truly differ.

The adaptation of an emerging virus to new hosts can affect the likelihood of future outbreaks and is a useful phenomenon to explore. Whilst computational evolutionary analysis can usefully indicate the presence of positive selection and molecular adaptation in virus genomes, sequence data alone cannot resolve the functional and biological context of selected sites. Further, different methods for detecting selection do not always give the same results, hence conclusions benefit from considering the

consensus of results generated by multiple models and methods (62). Three different models all indicated the positive selection acting on codon 66 of Gn (Table 1). The N-terminus of closely-related Bunyamwera virus (BUNV) Gn protein is predicted to localise on the outside of the viral particle (63) and function as a scaffold for the trimeric Gc spike (19). The residue encoded by codon 66 falls within this predicted extramembrane domain, potentially exposing it to the host immune system. Unlike other bunyaviruses, such as phleboviruses (64, 65) and hantaviruses (66, 67), it is unlikely that the Gn is large enough to extend from the virion membrane and shield the fusion loops of the cognate Gc. Thus, while amino acid changes are unlikely to affect fusion loop shielding, it is possible that selective forces acting on Gn could just alter the stability of the Gn-Gc lattice during cell entry.

We also detected positive selection on codon 86 of the NSm. The function of this protein among the orthobunyaviruses is variable, ranging from being essential for viral assembly in BUNV (particularly the N-terminal hydrophobic domains) (68) to being non-essential for viral growth in OROV (7) and other related pathogens such as Maguari virus (MAGV) (69). In Rift Valley fever virus (RVFV), NSm is non-essential for viral replication (70, 71) but plays a role in the suppression of apoptosis in infected cells (72, 73). Other potential roles for NSm in different bunyaviruses have been proposed (74), including a role in the infection of arthropod vectors (75). The reason for the adaptive evolution in OROV NSm detected here is therefore unclear.

The N-terminal head and stem domains of the orthobunyaviral Gc spike are the only regions for which atomic structures are available (20), making structure-based mapping

of sites feasible. While little is known about the host immune response to OROV, SBV Gc has been shown to be a target of antibody-mediated immune responses following infection (76), and is therefore likely to be under host selective pressure. Codons 269 and 442 (identified as evolving under positive selection) encode residues located in different domains of the Gc protein, in the first stem domain and core region, respectively. We mapped codon 269 to a loop at the N-terminus of the stem domain I, which connects to the head domain. Although we cannot discount a potential role in shielding the fusion loop in the core region, crystallographic analysis and comparison of this region with other reported orthobunyavirus Gc structures suggests that this loop may act as a flexible hinge that permits fusogenic rearrangements of the spike complex (20). Furthermore, and in line with the hypothesis that this region is important in stabilizing the mature spike complex, an identified neutralising antibody against SBV directly interacts with residues closer to the C-terminus of the protein's head domain, potentially disrupting Gc-mediated trimerization of the spike protein, as presented on viral particles (20). A similar immunogenic mechanism could play a role in the evolution of residue G269, rendering it a potential target of the neutralising antibody response. Of the residues that were identified as evolving under diversifying selection by the MEME approach alone, codon 145 encodes a residue that is solvent accessible and mapped to an alpha helix in the head domain (20), specifically the tip of the trimeric spike complex. This latter site could also constitute an antibody epitope, even if the evidence for this is presently weak (only one method identifies this residue as evolving under positive selection). Furthermore, if this region includes an epitope, the acquisition of N-linked glycosylation may mask the protein surface, similar to the "glycan shields" described

for Old World arenaviruses and HIV-1 (77, 78). N-linked glycosylation that is not essential for replication but increases virus infectivity of new cells has been described in BUNV (79) and RVFV (80), suggesting that different bunyaviruses could gain and lose glycosylation sites as they evolve, as a mechanism to mask spike epitopes from the humoral immune system. Mechanisms other than immune selection might play a role. For example, mutations in the E1 protein of Chikungunya virus affect its infectivity in different vector species (A226V increases infectivity in *Aedes albopictus* compared to *Aedes aegypti*) (81). The geographical spread of our isolates means that OROV variation could be associated with different host or vector species in South America; however, current data is insufficient to explore this idea further.

Our report incorporates new OROV genomes obtained from febrile patients in the province of Esmeraldas in northern Ecuador, representing the first direct detection and whole genome sequencing of the virus in the country (14). Previous prospective studies have suggested that OROV may circulate in both the coastal and Amazonian provinces of Ecuador (18, 82), but failed to directly detect or isolate the virus. Our sequences probably represent an undetected outbreak of OROV on Ecuador's northern coast, which may also include the southern coast of Colombia. Our phylogenetic analyses show that the Ecuadorian isolates form a single lineage that is closely related to viruses circulating in Peru, indicating a contiguous expansion of the known area of OROV circulation. The Ecuadorian isolates carry a distinct M segment that is shared with sequences from Peru and Panama, suggesting the presence of a distinct OROV lineage in the western, Pacific-bordering countries of South America.

OROV continues to be a neglected tropical disease and the numbers of genetic sequences for the virus are still limited. The genetic data that are available tend to be biased towards the shorter S segment, which has lower phylogenetic resolution and provides only a partial picture of OROV evolution. Available sequences are for the most part limited to viral diversity detected during human outbreaks, and therefore fail to capture the dynamics of OROV in reservoir host species and vector populations. The generation of whole OROV genomes – for example, generated locally using new rapid and portable sequencing technologies (83, 84); and the broadening of sampling to other countries and host species, are necessary to improve our understanding of the extent and nature of OROV transmission and the importance of reassortment for the evolution of this pathogen.

References

1. Hart TJ, Kohl A, Elliott RM. 2009. Role of the NSs protein in the zoonotic capacity of Orthobunyaviruses. Zoonoses Public Health 56:285-96.
2. Elliott RM. 2014. Orthobunyaviruses: recent genetic and structural insights. Nat Rev Microbiol 12:673-85.
3. Travassos da Rosa JF, de Souza WM, Pinheiro FP, Figueiredo ML, Cardoso JF, Acrani GO, Nunes MRT. 2017. Oropouche Virus: Clinical, Epidemiological, and Molecular Aspects of a Neglected Orthobunyavirus. Am J Trop Med Hyg 96:1019-1030.
4. Rodriguez-Morales AJ, Paniz-Mondolfi AE, Villamil-Gomez WE, Navarro JC. 2017. Mayaro, Oropouche and Venezuelan Equine Encephalitis viruses: Following in the footsteps of Zika? Travel Med Infect Dis 15:72-73.
5. Anderson CR, Spence L, Downs WG, Aitken TH. 1961. Oropouche virus: a new human disease agent from Trinidad, West Indies. Am J Trop Med Hyg 10:574-8.
6. Azevedo RS, Nunes MR, Chiang JO, Bensabath G, Vasconcelos HB, Pinto AY, Martins LC, Monteiro HA, Rodrigues SG, Vasconcelos PF. 2007. Reemergence of Oropouche fever, northern Brazil. Emerg Infect Dis 13:912-5.
7. Tilston-Lunel NL, Hughes J, Acrani GO, da Silva DE, Azevedo RS, Rodrigues SG, Vasconcelos PF, Nunes MR, Elliott RM. 2015. Genetic analysis of

members of the species Oropouche virus and identification of a novel M segment sequence. J Gen Virol 96:1636-50.
8. Vasconcelos HB, Azevedo RS, Casseb SM, Nunes-Neto JP, Chiang JO, Cantuaria PC, Segura MN, Martins LC, Monteiro HA, Rodrigues SG, Nunes MR, Vasconcelos PF. 2009. Oropouche fever epidemic in Northern Brazil: epidemiology and molecular characterization of isolates. J Clin Virol 44:129-33.
9. Cardoso BF, Serra OP, Heinen LB, Zuchi N, Souza VC, Naveca FG, Santos MA, Slhessarenko RD. 2015. Detection of Oropouche virus segment S in patients and inCulex quinquefasciatus in the state of Mato Grosso, Brazil. Mem Inst Oswaldo Cruz 110:745-54.
10. da Costa VG, de Rezende Feres VC, Saivish MV, de Lima Gimaque JB, Moreli ML. 2017. Silent emergence of Mayaro and Oropouche viruses in humans in Central Brazil. Int J Infect Dis 62:84-85.
11. Alvarez-Falconi PP, Rios Ruiz BA. 2010. [Oropuche fever outbreak in Bagazan, San Martin, Peru: epidemiological evaluation, gastrointestinal and hemorrhagic manifestations]. Rev Gastroenterol Peru 30:334-40.
12. Romero-Alvarez D, Escobar LE. 2017. Vegetation loss and the 2016 Oropouche fever outbreak in Peru. Mem Inst Oswaldo Cruz 112:292-298.
13. Watts DM, Phillips I, Callahan JD, Griebenow W, Hyams KC, Hayes CG. 1997. Oropouche virus transmission in the Amazon River basin of Peru. Am J Trop Med Hyg 56:148-52.
14. Wise EL, Pullan ST, Marquez S, Paz V, Mosquera JD, Zapata S, Jackson SK, Fejer G, Trueba G, Logue CH. 2018. Isolation of Oropouche Virus from Febrile Patient, Ecuador. Emerg Infect Dis 24:935-937.
15. Nunes MR, Martins LC, Rodrigues SG, Chiang JO, Azevedo Rdo S, da Rosa AP, Vasconcelos PF. 2005. Oropouche virus isolation, southeast Brazil. Emerg Infect Dis 11:1610-3.
16. Romero-Alvarez D, Escobar LE. 2018. Oropouche fever, an emergent disease from the Americas. Microbes Infect 20:135-146.
17. Sakkas H, Bozidis P, Franks A, Papadopoulou C. 2018. Oropouche Fever: A Review. Viruses 10.
18. Manock SR, Jacobsen KH, de Bravo NB, Russell KL, Negrete M, Olson JG, Sanchez JL, Blair PJ, Smalligan RD, Quist BK, Espin JF, Espinoza WR, MacCormick F, Fleming LC, Kochel T. 2009. Etiology of acute undifferentiated febrile illness in the Amazon basin of Ecuador. Am J Trop Med Hyg 81:146-51.
19. Bowden TA, Bitto D, McLees A, Yeromonahos C, Elliott RM, Huiskonen JT. 2013. Orthobunyavirus ultrastructure and the curious tripodal glycoprotein spike. PLoS Pathog 9:e1003374.
20. Hellert J, Aebischer A, Wernike K, Haouz A, Brocchi E, Reiche S, Guardado-Calvo P, Beer M, Rey FA. 2019. Orthobunyavirus spike architecture and recognition by neutralizing antibodies. Nat Commun 10:879.

21. Briese T, Bird B, Kapoor V, Nichol ST, Lipkin WI. 2006. Batai and Ngari viruses: M segment reassortment and association with severe febrile disease outbreaks in East Africa. J Virol 80:5627-30.
22. Saeed MF, Wang H, Suderman M, Beasley DW, Travassos da Rosa A, Li L, Shope RE, Tesh RB, Barrett AD. 2001. Jatobal virus is a reassortant containing the small RNA of Oropouche virus. Virus Res 77:25-30.
23. Aguilar PV, Barrett AD, Saeed MF, Watts DM, Russell K, Guevara C, Ampuero JS, Suarez L, Cespedes M, Montgomery JM, Halsey ES, Kochel TJ. 2011. Iquitos virus: a novel reassortant Orthobunyavirus associated with human illness in Peru. PLoS Negl Trop Dis 5:e1315.
24. Navarro JC, Giambalvo D, Hernandez R, Auguste AJ, Tesh RB, Weaver SC, Montanez H, Liria J, Lima A, Travassos da Rosa JF, da Silva SP, Vasconcelos JM, Oliveira R, Vianez JL, Jr., Nunes MR. 2016. Isolation of Madre de Dios Virus (Orthobunyavirus; Bunyaviridae), an Oropouche Virus Species Reassortant, from a Monkey in Venezuela. Am J Trop Med Hyg 95:328-38.
25. Bernardes-Terzian AC, de-Moraes-Bronzoni RV, Drumond BP, Da Silva-Nunes M, daSilva NS, Urbano-Ferreira M, Speranca MA, Nogueira ML. 2009. Sporadic oropouche virus infection, acre, Brazil. Emerg Infect Dis 15:348-50.
26. Vasconcelos HB, Nunes MR, Casseb LM, Carvalho VL, Pinto da Silva EV, Silva M, Casseb SM, Vasconcelos PF. 2011. Molecular epidemiology of Oropouche virus, Brazil. Emerg Infect Dis 17:800-6.
27. Wise EL, Márquez S, Mellors J, Paz V, Atkinson B, Gutierrez B, Zapata S, Coloma J, Pybus OG, Jackson SK, Trueba G, Fejer G, Logue CH, Pullan ST. 2019. Oropouche virus cases identified in Ecuador using an optimised rRT-PCR informed by metagenomic sequencing. bioRxiv.
28. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM, Dodd R, Mulembakani P, Schneider BS, Muyembe-Tamfum JJ, Stramer SL, Chiu CY. 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 7:99.
29. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-95.
30. Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. 2015. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks. PLoS One 10:e0143081.
31. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-9.
32. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 1:vev003.
33. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and postanalysis of large phylogenies. Bioinformatics 30:1312-3.

34. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2:vew007.
35. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016.
36. Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23:7-9.
37. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185-92.
38. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88.
39. Murray GG, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, Holmes MA, Rambaut A, Welch JJ. 2016. The effect of genetic structure on molecular dating and tests for temporal signal. Methods Ecol Evol 7:80-89.
40. Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-9.
41. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. 2018. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol doi:10.1093/molbev/msx335.
42. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Kosakovsky Pond SL. 2015. Gene-wide identification of episodic selection. Mol Biol Evol 32:1365-71.
43. Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-22.
44. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764.
45. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32:1342-53.
46. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol 30:1196-205.
47. Schrodinger, LLC. 2015. The PyMOL Molecular Graphics System, Version 1.8.
48. Krissinel E, Henrick K. 2007. Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774-97.
49. Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res 42:W320-4.

50. Lemey P, Minin VN, Bielejec F, Kosakovsky Pond SL, Suchard MA. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. Bioinformatics 28:3248-56.
51. Argimon S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093.
52. Nunes MRT, de Souza WM, Savji N, Figueiredo ML, Cardoso JF, da Silva SP, da Silva de Lima CP, Vasconcelos HB, Rodrigues SG, Ian Lipkin W, Vasconcelos PFC, Palacios G. 2019. Oropouche orthobunyavirus: Genetic characterization of full-length genomes and development of molecular methods to discriminate natural reassortments. Infect Genet Evol 68:16-22.
53. Burton DR. 2002. Antibodies, viruses and vaccines. Nat Rev Immunol 2:706-13.
54. Harrison SC. 2015. Viral membrane fusion. Virology 479-480:498-507.
55. Iroegbu CU, Pringle CR. 1981. Genetic interactions among viruses of the Bunyamwera complex. J Virol 37:383-94.
56. Urquidi V, Bishop DH. 1992. Non-random reassortment between the tripartite RNA genomes of La Crosse and snowshoe hare viruses. J Gen Virol 73 (Pt 9):2255-65.
57. Briese T, Calisher CH, Higgs S. 2013. Viruses of the family Bunyaviridae: are all available isolates reassortants? Virology 446:207-16.
58. Wohl S, Schaffner SF, Sabeti PC. 2016. Genomic Analysis of Viral Outbreaks. Annu Rev Virol 3:173-195.

59. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. Trends in Ecology & Evolution 18:481-488.
60. Freire CC, Iamarino A, Soumare PO, Faye O, Sall AA, Zanotto PM. 2015. Reassortment and distinct evolutionary dynamics of Rift Valley Fever virus genomic segments. Sci Rep 5:11353.
61. Lam TT, Liu W, Bowden TA, Cui N, Zhuang L, Liu K, Zhang YY, Cao WC, Pybus OG. 2013. Evolutionary and molecular analysis of the emergent severe fever with thrombocytopenia syndrome virus. Epidemics 5:1-10.
62. Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950-8.
63. Shi X, Botting CH, Li P, Niglas M, Brennan B, Shirran SL, Szemiel AM, Elliott RM. 2016. Bunyamwera orthobunyavirus glycoprotein precursor is processed by cellular signal peptidase and signal peptide peptidase. Proc Natl Acad Sci U S A 113:8825-30.
64. Allen ER, Krumm SA, Raghwani J, Halldorsson S, Elliott A, Graham VA, Koudriakova E, Harlos K, Wright D, Warimwe GM, Brennan B, Huiskonen JT, Dowall SD, Elliott RM, Pybus OG, Burton DR, Hewson R, Doores KJ, Bowden TA. 2018. A Protective Monoclonal Antibody Targets a Site of

Vulnerability on the Surface of Rift Valley Fever Virus. Cell Rep 25:3750-3758 e4.
65. Halldorsson S, Li S, Li M, Harlos K, Bowden TA, Huiskonen JT. 2018. Shielding and activation of a viral membrane fusion protein. Nat Commun 9:349.
66. Li S, Rissanen I, Zeltina A, Hepojoki J, Raghwani J, Harlos K, Pybus OG, Huiskonen JT, Bowden TA. 2016. A Molecular-Level Account of the Antigenic Hantaviral Surface. Cell Rep 16:278.
67. Rissanen I, Stass R, Zeltina A, Li S, Hepojoki J, Harlos K, Gilbert RJC, Huiskonen JT, Bowden TA. 2017. Structural Transitions of the Conserved and Metastable Hantaviral Glycoprotein Envelope. J Virol 91.
68. Shi X, Kohl A, Leonard VH, Li P, McLees A, Elliott RM. 2006. Requirement of the Nterminal region of orthobunyavirus nonstructural protein NSm for virus assembly and morphogenesis. J Virol 80:8089-99.
69. Pollitt E, Zhao J, Muscat P, Elliott RM. 2006. Characterization of Maguari orthobunyavirus mutants suggests the nonstructural protein NSm is not essential for growth in tissue culture. Virology 348:224-32.
70. Gerrard SR, Bird BH, Albarino CG, Nichol ST. 2007. The NSm proteins of Rift Valley fever virus are dispensable for maturation, replication and infection. Virology 359:459-65.
71. Won S, Ikegami T, Peters CJ, Makino S. 2006. NSm and 78-kilodalton proteins of Rift Valley fever virus are nonessential for viral replication in cell culture. J Virol 80:82748.
72. Won S, Ikegami T, Peters CJ, Makino S. 2007. NSm protein of Rift Valley fever virus suppresses virus-induced apoptosis. J Virol 81:13335-45.
73. Terasaki K, Won S, Makino S. 2013. The C-terminal region of Rift Valley fever virus NSm protein targets the protein to the mitochondrial outer membrane and exerts antiapoptotic function. J Virol 87:676-82.
74. Eifan S, Schnettler E, Dietrich I, Kohl A, Blomstrom AL. 2013. Non-structural proteins of arthropod-borne bunyaviruses: roles and functions. Viruses 5:2447-68.
75. Crabtree MB, Kent Crockett RJ, Bird BH, Nichol ST, Erickson BR, Biggerstaff BJ, Horiuchi K, Miller BR. 2012. Infection and transmission of Rift Valley fever viruses lacking the NSs and/or NSm genes in mosquitoes: potential role for NSm in mosquito infection. PLoS Negl Trop Dis 6:e1639.
76. Wernike K, Aebischer A, Roman-Sosa G, Beer M. 2017. The N-terminal domain of Schmallenberg virus envelope protein Gc is highly immunogenic and can provide protection from infection. Sci Rep 7:42500.
77. Watanabe Y, Raghwani J, Allen JD, Seabright GE, Li S, Moser F, Huiskonen JT, Strecker T, Bowden TA, Crispin M. 2018. Structure of the Lassa virus glycan shield provides a model for immunological resistance. Proc Natl Acad Sci U S A 115:7320-7325.
78. Seabright GE, Doores KJ, Burton DR, Crispin M. 2019. Protein and Glycan Mimicry in HIV Vaccine Design. J Mol Biol 431:2223-2247.

79. Shi X, Brauburger K, Elliott RM. 2005. Role of N-linked glycans on bunyamwera virus glycoproteins in intracellular trafficking, protein folding, and virus infectivity. J Virol 79:13725-34.
80. Phoenix I, Nishiyama S, Lokugamage N, Hill TE, Huante MB, Slack OA, Carpio VH, Freiberg AN, Ikegami T. 2016. N-Glycans on the Rift Valley Fever Virus Envelope Glycoproteins Gn and Gc Redundantly Support Viral Infection via DC-SIGN. Viruses 8.
81. Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. 2007. A single mutation in chikungunya virus affects vector specificity and epidemic potential. PLoS Pathog 3:e201.
82. Forshey BM, Guevara C, Laguna-Torres VA, Cespedes M, Vargas J, Gianella A, Vallejo E, Madrid C, Aguayo N, Gotuzzo E, Suarez V, Morales AM, Beingolea L, Reyes N, Perez J, Negrete M, Rocha C, Morrison AC, Russell KL, Blair PJ, Olson JG, Kochel TJ, Group NFSW. 2010. Arboviral etiologies of acute febrile illnesses in Western South America, 2000-2007. PLoS Negl Trop Dis 4:e787.
83. Faria NR, Quick J, Claro IM, Theze J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, Franco LC, Silva SP, Wu CH, Raghwani J, Cauchemez S, du Plessis L, Verotti MP, de Oliveira WK, Carmo EH, Coelho GE, Santelli A, Vinhal LC, Henriques CM, Simpson JT, Loose M, Andersen KG, Grubaugh ND, Somasekar S, Chiu CY, Munoz-Medina JE, Gonzalez-Bonilla CR, Arias CF, Lewis-Ximenez LL, Baylis SA, Chieppe AO, Aguiar SF, Fernandes CA, Lemos PS, Nascimento BLS, Monteiro HAO, Siqueira IC, de Queiroz MG, de Souza TR, Bezerra JF, Lemos MR, Pereira GF, Loudal D, Moura LC, Dhalia R, Franca RF, et al. 2017. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 546:406-410.
84. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG, Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Alcantara LC, Jr., Sabino EC, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12:1261-1276.

Tables

Table 1. Evidence for adaptive selection on the three genome segments of the OROV genome assessed using different models.

Sites identified by more than one method are highlighted in bold.

			Pervasive Selection								Episodic Selection			
			Gene-wise (BUSTED)		*Site-wise (SLAC)*		*Site-wise (FEL)*		*Site-wise (FUBAR)*		*Site-wise (MEME)*		*Branch-site model (aBSREL)*	
Segment	Gene	Length	*p* <0.05	Mean *dN/dS*[a]	*p* <0.05	Sites	*p* <0.05	Sites	Test <0.05	Sites	*p* <0.05	Sites	*p* <0.05	Sites
S	N/Ns	693 nt	YES	0.071	NO	--	NO	--	YES	1089	YES	2	NO	--
M	Gn	918 nt	NO	0.055	YES	**66**	YES	**66**	YES	**66**, 274	YES	**66**	NO	--
	NSm	525 nt	NO	0.114	NO	--	YES	**86**	YES	**86**	YES	**86**	NO	--
	Gc	2817 nt	NO	0.050	NO	--	YES	**269**, **442**	YES	**269**, **442**	YES	71, 106, 145, **269**, **442**, 645, 842, 879	NO	--
L	RdRP	6756 nt	NO	0.027	NO	--	NO	--	YES	230	YES	237, 1089, 1394, 1397, 2048, 2085	NO	--

[a] Estimated as weighed ω mean values from the unconstrained model.

3 Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador

Highlight: A severe, *undersampled* first wave leaves traces on the inferred transmission lineages

This phylogenetic analysis of the first year of SARS-CoV-2 circulation in Ecuador was performed with a relatively small genomic data set by current standards. By comparing epidemiological and genomic data we show that the severe first wave of the epidemic in the country was very sparsely represented in the genomic data. However, when inferring instances of viral importation from international locations into Ecuador, we can identify the signal of this limited sampling during the early stages of the epidemic in the form of long basal branches in the phylogenetic trees. It is also evident that the viral imports had widespread circulation within the country, and that the dynamics and composition of lineages between the two largest cities in the country are considerably different. This points towards the possibility of different importation dynamics across regions, which can be hypothesised despite the small data set size.

Abstract

Characterisation of SARS-CoV-2 genetic diversity through space and time can reveal trends in virus importation and domestic circulation, and permit the exploration of questions regarding the early transmission dynamics. Here we present a detailed description of SARS-CoV-2 genomic epidemiology in Ecuador, one of the hardest hit countries during the early stages of the COVID-19 pandemic. We generated and analysed 160 whole genome sequences sampled from all provinces of Ecuador in 2020.

Molecular clock and phylogeographic analysis of these sequences in the context of global SARS-CoV-2 diversity enable us to identify and characterise individual transmission lineages within Ecuador, explore their spatiotemporal distributions, and consider their introduction and domestic circulation. Our results reveal a pattern of multiple international importations across the country, with apparent differences between key provinces. Transmission lineages were mostly introduced before the implementation of non-pharmaceutical interventions (NPIs), with differential degrees of persistence and national dissemination.

Introduction

The rapid generation of substantial numbers of virus genomic sequences during the COVID-19 pandemic is without precedent. Laboratories and institutes around the world produced and shared over 300,000 whole SARS-CoV-2 genome sequences in the GISAID repository during 2020, and features a total of 1,428,296 sequence entries by May 10, 2021 (*1*), providing an unparalleled data set that permits detailed analyses of virus transmission and dissemination. These achievements have provided insights into the sources of SARS-CoV-2 importation and early transmission dynamics in individual countries and geographical regions (*2*-*5*), and have enabled the exploration of viral transmission history at a global scale (*6*). Phylogenetic methods, including molecular clock models and phylogeographic and phylodynamic methods, are now used routinely to analyse such genomic data from emerging outbreaks (*7*). The resolution level of evolutionary and transmission history obtained using these methods is contingent on the virus' evolutionary rate and the depth and representativeness of sampling of cases across

space and time (*8*). While heterogeneous sampling and sequencing among countries can bias and affect the output of some phylogeographic methods (*9, 10*), general trends in the transmission of viral lineages can still be inferred from smaller samples of genomic sequences from individual locations.

The utility of pathogen genomic surveillance during outbreaks has developed during various past emerging epidemics (*11-13*) and has gained further momentum during the current global health crisis. Information about epidemiological trends can be effectively complemented with genomic analyses in order to understand case-specific (*14*) and general transmission patterns (*15*). This framework can be extended to account for other factors that affect the spread of pathogens, ranging from human mobility on a global scale (*16*) to particular social networks (*17*). The analysis of local- and national-scale data sets during the current pandemic has provided insights into the processes affecting the introduction and circulation of the virus into new locations (*18*) and provided genomic context for other data sources (*19-21*). Indeed, the integration and analysis of multiple data sources about an emerging epidemic has the potential to compensate for surveillance blind spots and better understand poorly sampled outbreaks (*22*).

The COVID-19 epidemic in Ecuador was marked by a dramatic and widely publicised early phase (*23*) with an estimated basic reproductive number (R_0) of 3.54 (*24*). Ecuador is a small middle-income South American country with the seventh largest population in the continent; half of the country's population lives in Guayas province (host of the country's most populated city, Guayaquil) and Pichincha province (host of the country's second most populated city, the capital Quito; Fig 1A). The first case was

reported in the country on 27th February 2020 (a patient who returned from abroad through Guayaquil with date of symptom onset of February 15) and was followed by the declaration of a National Health Emergency on 11th March 2020. Public health interventions were implemented shortly thereafter: mass gatherings were restricted on March 13, and a partial lockdown that included the closure of international borders was implemented on March 17. Finally, a full lockdown that included a curfew and the limitation of domestic mobility in private and public vehicles came into effect on March 25 (*24*). The country's port city of Guayaquil was the first epicentre of the epidemic, facing a severe increase in the numbers of cases between late February and early April. The province of Guayas reached its highest effective reproductive number R_t (defined as the average number of secondary cases caused by a primary cases at a point in time t; (*25*) on March 14 (R_t estimates vary between 3.96 and 4.91) with 1462 cases reported that day; (*26*) and reported a cumulative incidence of 146.94 cases per 100000 people by April 18 (*24*). The actual number of cases are likely to have been much higher when evaluated through the lens of excess mortality data (as obtained from the National Institute of Statistics and Census; Instituto Nacional de Estadística y Censos, INEC), and would explain why local diagnostic and healthcare services became rapidly overwhelmed (*23, 24*). After the peak and decline of the epidemic's first wave in Ecuador, restrictions were maintained during April and May and progressively relaxed over the following months, as the epicentre of Ecuador's epidemic moved to the capital city of Quito, located in Pichincha province (which on 23rd July 2020 overtook Guayaquil as the city with the greatest number of COVID-19 confirmed cases). The last

restrictions were finally lifted on September 13, although use of personal protective equipment and physical distancing guidelines remained in place.

To date, the source and diversity of circulating transmission lineages in Ecuador and their reach across the country remain unexplored. International importations are expected to have played an important role in seeding transmission chains in Ecuador, as observed in other countries in Latin America (*5, 27, 28*) and elsewhere (*15*). We undertake phylogenetic analyses of 160 SARS-CoV-2 whole genome sequences sampled from Ecuadorian cases and place them within the context of global viral genetic diversity in order to characterise the genomic epidemiology of SARS-CoV-2 in the country. We identify introduction events and transmission lineages within Ecuador and investigate their spatio-temporal distribution, and we hypothesise about the role of domestic seeding on viral transmission dynamics.

Methods

Genomic sequencing of SARS-CoV-2 samples from Ecuador

Clinical samples were collected from patients with a laboratory confirmed SARS-CoV-2 infection in Ecuador during 2020 (Fig 1A-C). Samples collected by the Microbiology Institute at Universidad San Francisco de Quito (IM-USFQ) were obtained from third-level hospitals (i.e. specialised tertiary referral hospitals) across all 24 provinces in the country without unified selection criteria (*29*). Samples collected by the Omics Sciences Laboratory at Universidad de Especialidades Espiritu Santo (UEES) were obtained from samples collected by the laboratory's diagnostic service and selected

at random for sequencing. Samples collected by the National Institute of Investigations in Public Health (Instituto Nacional de Investigación en Salud Pública, INSPI) were obtained from the national epidemiological SARS-CoV-2 surveillance system. Samples collected by the Biomedical Research Unit at Zurita & Zurita Laboratories (ZZL) were obtained from community patients from Quito who presented clinical signs of reinfection. This complete Ecuadorian sample set was collected between March 9 and December 9 2020, with limited representation during the early months of the epidemic when compared to excess mortality data (Fig 1C).

From these samples we generated 160 complete SARS-CoV-2 genomic sequences using different methodologies. From these, 121 genomes represent samples collected between the implementation of the national lockdown (March 25) and the lifting of restrictions (September 13). IM-USFQ generated 108 whole genome sequences using Oxford Nanopore MinION sequencing and the ARTIC Network primer scheme approach as previously described (*29*). UEES generated 33 whole genome sequences through Illumina sequencing on a MiniSeq platform (Illumina, San Diego, CA). INSPI generated 15 sequences either in collaboration with Charité - Berlin University of Medicine (through Illumina sequencing) or on site at the Centre for Multidisciplinary Research of the Direction of Research, Development and Innovation (through Oxford Nanopore MinION sequencing as described in *30*). Zurita & Zurita Laboratories generated 4 sequences through Illumina sequencing on a MiSeq platform (Illumina, San Diego, CA). Details of sample collection, sequencing and genome assembly are summarised in Table S1. Sample collection dates and the province-level geographical

location of residence of the patient were included as metadata for all sequences in the country.

To determine the viral genetic diversity circulating in the country during the sampling period, all sequences from Ecuador were phylogenetically assigned under the global Pango lineage system using the *Pangolin v2.2.2* tool (https://virological.org/t/pangolin-web-application-release/482).

Global SARS-CoV-2 data sets

The Ecuadorian virus sequences were analysed in the context of global SARS-CoV-2 genomic diversity by including all high-quality SARS-CoV-2 genome sequences and their accompanying metadata available in GISAID (*1*) on January 1 2021 (sequences were retained if they were > 29000 nucleotides long and < 5% of the sequence was missing). Sequences without a complete sample collection date or not attributed to human hosts were excluded, yielding a total of 218,771 sequences from samples collected from December 1, 2019 up until December 10, 2020.

The large number of SARS-CoV-2 genomes generated during 2020 makes full-scale phylogenomic analyses computationally prohibitive. We therefore subsampled sequences from the abovementioned full data set (*i.e.* all GISAID sequences included in our analyses, excluding the complete set of Ecuadorian sequences) using two approaches. First, we randomly sampled one sequence per country per day from the full data set over the complete sampling period, to create a "*systematically-subsampled data set"* (comprised of 8606 sequences). In parallel, we arbitrarily generated three

"*randomly-subsampled data sets"* consisting of 8606 randomly chosen sequences from the full data set, to match the size of the systematically-subsampled data set. These randomly-subsampled data sets were used to evaluate the performance of the background SARS-CoV-2 sequences as the genomic context for the identification of transmission lineages within Ecuador (see the *Phylogenetic identification of transmission lineages* section below), as an alternative to generating multiple bootstrap sampled of the real background sequence data which would require considerable longer analysis times. Finally, we added the sequences from Ecuador to each data set, resulting in a total of 8766 sequences per data set. We didn't weigh the subsampling by country of origin, resulting in an increased number of sequences from locations with improved genomic surveillance schemes predominantly in Europe and North America; however, given that these locations represented important epicentres during early and mid-2020 and that there is substantial connectivity between Ecuador and some of these locations (such as the USA and Spain), we expect that this subsampling would retain sufficient relevant sequences to inform new importations of the virus into the country.

Each data set was aligned to the Wuhan-Hu-1 (GenBank accession: MN908947.3) reference genome sequence (*31*) using *Minimap 2.17* (*32*) to generate multiple sequence alignments. Sites containing >90% gaps relative to the sequences in their respective alignment were masked, whilst the untranscribed terminal regions (UTRs) were trimmed. After masking and trimming, the resulting alignments had a final length of 29,409 nucleotides, with the shortest partial genome sequences being cut down to 28,955 nucleotides long.

Phylogenetic identification of transmission lineages

We followed a similar rationale and methodology to that described in du Plessis *et al* (*15*) to identify local transmission lineages. Phylogenetically linked sequences were inferred to have descended from a common ancestor if they were associated with a single inferred introduction event into Ecuador from an international location (*5, 15*). Ecuadorian transmission lineages therefore correspond to lineages of sequences sampled within the country that descend from a node inferred to have also occurred in Ecuador, which must in turn have descended from outside of the country. Given the unstructured sampling of the Ecuadorian sequences, some transmission lineages will likely correspond to epidemiologically linked cases (i.e. targeted investigation of epidemiological clusters); these have been identified as such in the text whenever the information was available.

Maximum likelihood (ML) phylogenetic trees were estimated from the systematically-subsampled data set and the randomly-subsampled data sets using *IQtree 2.1.1* (*33*) under a GTR+Γ substitution model. Node support was estimated through an SH-like approximate Likelihood Ratio Test (aLRT) using 1000 replicates for the SH-like algorithm that uses a resampling of estimated log-likelihoods (RELLs) bootstrap to estimate the confidence of the aLRT statistics (*34*). While the randomly-subsampled data sets were not analysed further, the tree for the systematically-subsampled data set was re-rooted by heuristically searching for the root placement that minimises the mean squared residual of a regression of sequence sample date against root-to-tip genetic

distance, calculated using *TempEst v1.5.3* (*35*), to maximise the temporal signal of the data set. The same regression was used to assess the clock-like behaviour of the data set.

Subsequent analyses in our pipeline require an evolutionary rate estimation. We performed an exploratory analysis on a random selection of 866 genomes from the systematically-subsampled data set (~10% of the sequences, following *15*, ensuring that representatives of the earliest and latest collection dates were included) in order to estimate the evolutionary rate of the data set over the sampling period with a computationally permissive data set. We used *BEAST v.1.10.4* (*36*) to obtain a clock rate estimate using the HKY substitution model and a strict molecular clock with a continuous-time Markov chain (CTMC) prior (*37*). We employed a Skygrid coalescent tree prior (*38*) that accounts for the 50 epidemiological weeks over which the genomes were sampled, plus a cut-off period that precedes the earliest collected SARS-CoV-2 sequences. Independent Markov Chain Monte Carlo (MCMC) chains were run for 40 million steps and subsequently combined after discarding the initial 10% of each run as a burn-in. Convergence of relevant parameters was assessed by visually inspecting the MCMC trace plots and posterior probability distributions of parameters from independent chains and using effective sample size (ESS) estimates approaching 100 from the combined chains using *Tracer v1.7.1* (*39*); these unusually low ESS values were chosen due to the complexity of some model parameters which resulted in suboptimal mixing of the MCMC.

The systematically subsampled data set was analysed with *BEAST v1.10.5* (https://github.com/beast-dev/beast-mcmc/releases/tag/v1.10.5pre_thorney) using a

newly implemented method that significantly reduces analysis time by using a simple model to estimate a time-calibrated tree (also called *Thorney BEAST*; see *40*). This approach takes a previously estimated rooted phylogenetic tree (henceforth called the *data tree*) instead of an alignment, and rescales branches in this tree into time. Under this model, the likelihood of each branch length (in mutations) is defined as a function of a Poisson distribution with a mean directly proportional to the clock rate (*40, 41*); we therefore used a rate of 6.28×10^{-4} substitutions/site/year, based on the median clock rate estimate obtained from our exploratory analysis. We defined a coalescent Skygrid prior, similar to the one described for the exploratory analysis, and used the previously mentioned re-rooted ML tree as a starting *data tree*. Independent MCMC chains were run for 100 million steps and combined (after discarding 10% of each run as burn-in) to produce a posterior tree distribution. Convergence was assessed through the examination of trace plots and ESS estimates approximating 100 as previously described.

To identify nodes associated with transmission lineages in Ecuador, we used a discrete phylogeographic model consisting of a two-state discrete trait analysis (DTA; *42*) implemented in *BEAST v1.10.4* (*36*). The posterior distribution of trees generated in the previous step was resampled down to 500 time-calibrated trees using *LogCombiner v1.10.4* and *BEAST* was used to sample this tree space. Tips were assigned to one of two possible states (Ecuador vs non-Ecuador) and reconstruction of ancestral node states was undertaken using an asymmetric substitution model (*42*). The expected number of DTA transitions between international locations and Ecuador were estimated using a robust

counting approach (*43*). Two independent MCMC chains of 5 million steps each were combined for this analysis, after discarding the first 500,000 steps of each run as burn-in. A Maximum Clade Credibility (MCC) tree was generated from the DTA posterior tree distribution by sampling 1000 trees from the combined MCMC runs in *TreeAnnotator v.1.10*. Each internal node was assigned a posterior probability for its inferred location, and these were used to evaluate uncertainty regarding the assignment of potential transmission lineages in Ecuador.

Transmission lineages and transmission lineage groups

All phylogenetic clusters of sequences from Ecuador were inspected visually on the MCC tree to assign individual transmission lineages. The nomenclature of these country-specific transmission lineages followed a one-letter code in alphabetical order, defined by the earliest sample collection date in each transmission lineage. General features of each transmission lineage were summarised, such as the earliest and latest sample in each lineage, the number of provinces in which the transmission lineage had been identified and the number of sequences belonging to said lineage (used as a proxy of transmission lineage size). The consistency with which sequences were grouped into these transmission lineages was evaluated by visually inspecting the ML trees for the randomly-subsampled data sets and comparing the clusters of Ecuadorian sequences to those from the DTA analysis.

Results

SARS-CoV-2 genetic diversity in Ecuador

The samples from laboratory-confirmed individuals obtained across mainland Ecuador (and one sample from the Galápagos Islands) were collected as they became available through different hospitals and laboratories and yielded genomes from all provinces in the country (Fig 1A). The number of sequences per province correlates with the cumulative number of excess deaths per province during 2020 (compared to the mean number of deaths per province between 2015 and 2019; Pearson's $R = 0.91$, $p =$ 6.9e-10; Spearman's *rho* = 0.556, $p = 0.005$), suggesting that the number of sequences per province is approximately proportional to the number of infections (Fig 1B). Despite the testing limitations in the country throughout 2020, the number of sequences also correlate to the cumulative number of patients with a positive or suspected positive COVID-19 PCR test over the sampling period (Pearson's $R = 0.56$, $p = 004$; Spearman's *rho* = 0.603, $p = 0.002$). Even with the limited number of sequences from Ecuador, the proportion of cases sequenced in our sample is similar to that of other countries in the region (*61*). We estimate Ecuador produced 8 sequences for every 10,000 reported cases or 12 sequences for every 1,000 officially reported COVID-19 deaths. This results in a higher coverage than Peru (4 sequences per 10,000 cases/10 sequences per 1,000 deaths) or Brazil (3 sequences per 10,000 cases/10 sequences per 1,000 deaths) but lower coverage than Uruguay (159 sequences per 10,000 cases, and a higher number of SARS-CoV-2 genome sequences than reported deaths) (Fig S1, File S1). It should be noted however that these estimates rely on the testing intensities between provinces in Ecuador

and between different countries; limited testing in Ecuador could mean that the overall representation is lower than estimable from official reports.

Wide variation in the total numbers of cases across provinces in Ecuador is reflected in variation in number of sequences obtained. While heavily affected provinces such as Pichincha (72,305 confirmed cases until December 10) and Guayas (26,080 confirmed cases) account for larger numbers of sequences (47 and 18 for Guayas and Pichincha respectively), less affected provinces in the southern Highlands (Azuay – 12,670 confirmed cases, and Loja – 7252) and the Amazon (Morona Santiago – 3422, Napo – 1605, Orellana – 2100, Pastaza – 2360 and Zamora Chinchipe – 1628) are represented by few sequences (28 in total). Only a single sequence was obtained from the Galápagos because of the extremely low number of cases in this province. Manabí province appears to be underrepresented (14,061 confirmed cases) while Imbabura (5695 confirmed cases) and Los Ríos (4707 confirmed cases) are represented by higher numbers of genomes per death (Fig. 1B). Sequence sampling rates for each province (excluding Galápagos) varied between 3 and 86 sequences per 1000 deaths (2). The temporal distribution of samples collected in Ecuador during 2020 do not strongly match trends in reported excess deaths. More samples were collected in July and August, but fewer genomes were sampled in the early epidemic months (March to May) despite the high number of excess deaths reported then (Fig 1C). Sequence coverage is greater for the coastal provinces during the early months of the epidemic (March to June), when the epicentre of the epidemic was based in the port city of Guayaquil (in the Guayas province, (*23*), and shifted towards higher sampling in the highlands and Amazon

provinces, as the epicentre of the epidemic shifted towards the capital city of Quito (in Pichincha province) and as more cases were reported in the Amazon.

Virus genomes from Ecuador were assigned to specific Pango lineages (*45*) using the *Pangolin* tool (https://virological.org/t/pangolin-web-application-release/482). The genomes were assigned to 33 different global lineages and predominantly B.1.1.74 (39.4% of all Ecuadorian sequences ; Fig. 1D), one of the lineages descended from B.1.1 which became one of the most dominant lineages during the early phase of the pandemic in Europe and North America (after the virus was introduced from Asia; (*3, 45*). The geographic distribution of the SARS-CoV-2 lineage diversity in the country shows distinctive patterns: the most prevalent Pango lineages (B.1.1.74, B.1 and B.1.1.1) are found on multiple provinces, while the majority of the lineages observed at low-frequencies were found in more heavily affected (and therefore better sampled) provinces, particularly Guayas and Pichincha. The heavily affected (and highly populated) province of Guayas (where Guayaquil is located) exhibits a predominance of the B.1.1.74 lineage (59.1% of all sequences from this province; Fig 1D). B.1.1.74 is also abundant in the provinces of Los Ríos, which neighbours Guayas (49%), and Imbabura, which neighbours Pichincha (38.9%). Other common lineages, such as B.1 (16.9% of all sequences from Ecuador) and B.1.1.1 (6.9% of all sequences from Ecuador), are distributed across various geographical regions.

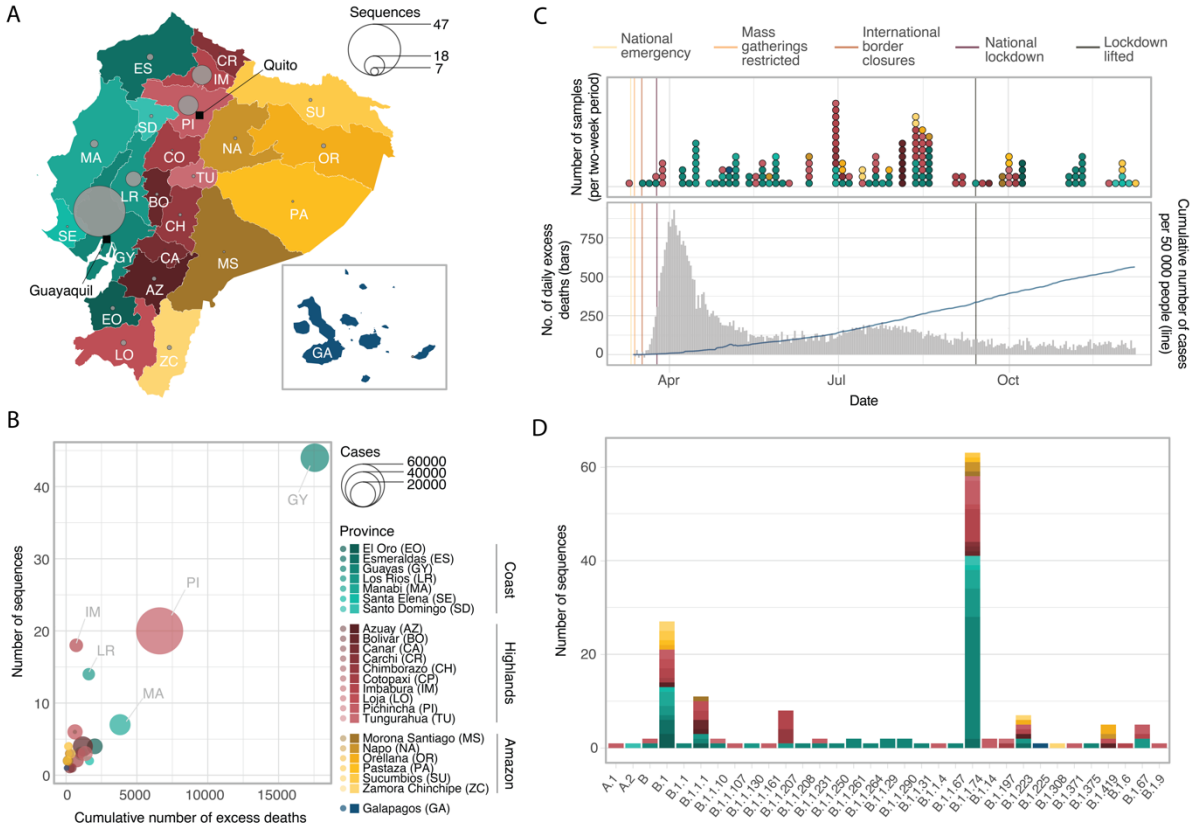


Figure 1. Overview of genomic sampling and SARS-CoV-2 genetic diversity in Ecuador.

A. Number of sequences from Ecuador analysed in this study per province across the four main geographic regions: the coast (shades of green), the highlands (shades of yellow), the Amazon (shades of orange) and the Galápagos (blue). ***B.*** Number of deaths (attributed to laboratory-confirmed COVID-19 cases) versus number of whole genome sequences available per province. Circle radius shows the number of cases per province. ***C.*** Timelines showing collection dates of sequences from the four geographic regions across time (upper panel) and the COVID-19 epidemiological curves in Ecuador during 2020 (cumulative number of laboratory-confirmed cases as reported by the Ministry of Health in the blue line, number of daily excess deaths compared to the same dates in 2019 as reported by the National Institute of Statistics and Census in grey; lower panel). **D.** Geographic distribution of SARS-CoV-2 lineages identified in Ecuador.

Identification of Ecuadorian transmission lineages

We estimated maximum likelihood (ML) trees of the Ecuadorian sequences in the context of different background data sets and performed Bayesian phylogenetic

inference on a systematically-subsampled data set. The clustering patterns of Ecuadorian sequences in the ML trees showed some variation between data sets, but in the majority of cases remained consistent (File S3). We therefore derive our results from the systematically subsampled data set and discuss these in light of the randomly subsampled data sets.

We consistently found that a sizeable proportion of sequences from Ecuador do not cluster with other sequences from the country (54/160 sequences for the systematically subsampled data set, 48 to 51/160 for the randomly-subsampled data sets; File S3), and were therefore assigned as singletons and not associated with further virus spread within Ecuador detectable through genomic analysis. These singletons could in fact represent introduction events that produced no forward transmission, or cases where forward transmission did occur but was not captured in this study due to the small sample size. We note that the majority of the singleton sequences were collected before mid-July (Fig S2); we speculate that they could represent predominantly early introduction events that occurred before the implementation of a national lockdown on March 16. While it is possible to establish a possible limit of dates on which each singleton was introduced, based on the last ancestral node inferred to have occurred outside of Ecuador, the precise importation date will fall somewhere between the inferred age of this preceding node and the collection date of the singleton sequence. The low sampling density in Ecuador and our subsampling schemes are likely to introduce uncertainty in estimating the age of these nodes and we therefore excluded these analyses from our results.

The remaining sequences (106/160 sequences for the systematically-subsampled data set) fall into two distinct categories. Firstly, 20 monophyletic clusters of Ecuadorian sequences were identified, capturing multiple introduction events and some local viral circulation patterns. These clusters were assigned to be separate Ecuadorian transmission lineages, named A through V (with exceptions detailed in the paragraph below; Fig 2A, Fig S3-S9). Each represents a single introduction event of the virus from an international destination, followed by local forward transmission in Ecuador (*15*).

Secondly, we identified two large monophyletic clusters that include sequences from international locations and Ecuador. These were not identified strictly as individual transmission lineages through our DTA approach, but rather as genetically similar groups of individual transmission lineages. While these may represent clusters of independent introductions from shared sources, there is also a possibility that these groups of transmission lineages in fact correspond to single introduction events misidentified by our DTA analysis, given the variation in intensity of SARS-CoV-2 sampling across countries (*6, 9*), including Ecuador, and the limited genetic divergence observed in SARS-CoV-2 over the time span being analysed (*44*). This likely resulted in the ancestral nodes being inferred to have existed outside of Ecuador due to the phylogenetic placement of the Ecuadorian sequences (Fig 2A). While there is a possibility that these in fact represent multiple closely related yet independently introduced transmission lineages, we here identify them as transmission lineage groups (labelled with a single letter and highlighted with an asterisk, D* and H*) for summary

purposes. The remaining unambiguous transmission lineages were each identified with their own letters and are shown in Figures S3-S9.

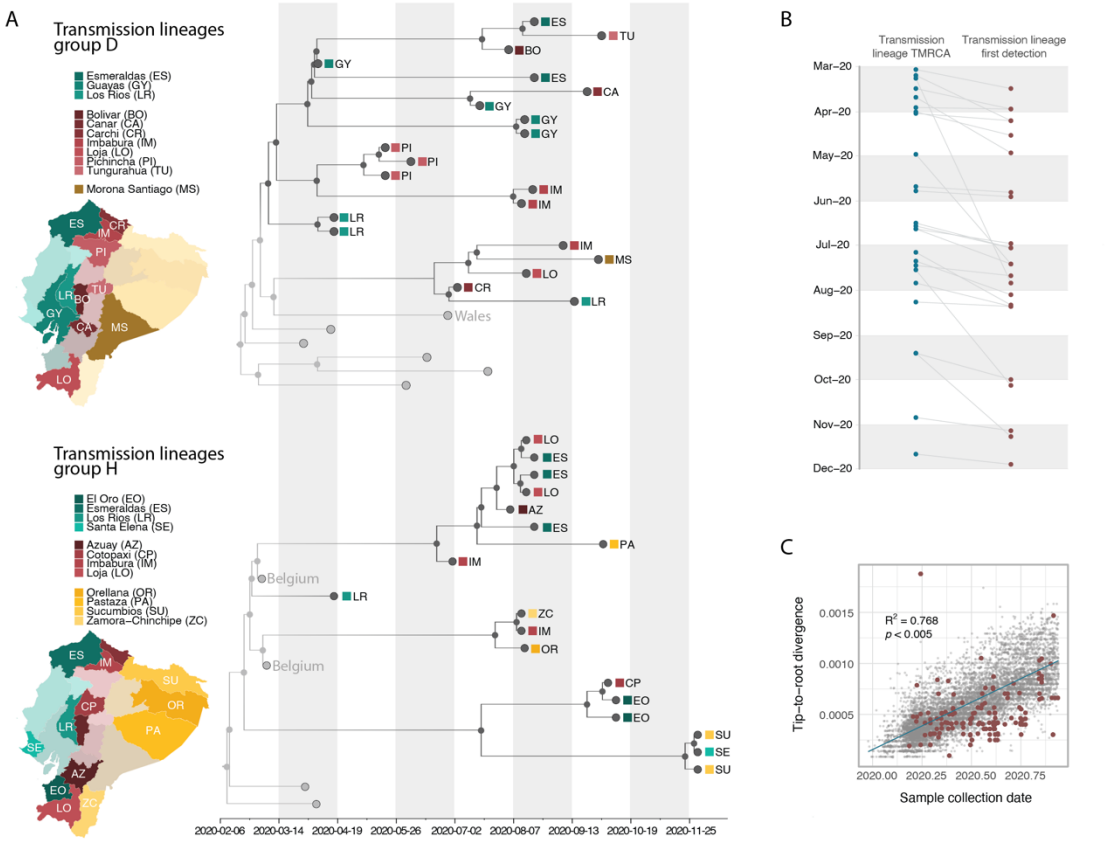


Figure 2. Time calibrated phylogenetic trees for the major transmission lineages in Ecuador.

A. Subtrees extracted from a time-calibrated Maximum Clade Credibility (MCC) tree of SARS-CoV-2 whole genome sequences, corresponding to the two largest clusters of sequences from Ecuador. Tree tips are coloured by sampling location (in Ecuador, red, versus outside of Ecuador, grey); nodes and branches are coloured by inferred location through a two-state DTA analysis. The province where each sequence was sampled is annotated on the tips, and maps highlight these provinces. Tips that correspond to sequences that cluster together within the major Ecuadorian clusters are also annotated with the region where the samples were collected. ***B.*** Detection lag of individual transmission lineages in Ecuador, showing the median TMRCA of each transmission lineage from our data set (blue) connected by a grey line to the date of the earliest sequence in that transmission lineage (red). ***C.*** Root -to-tip genetic distances (based on

a heuristically rooted Maximum Likelihood tree) versus sample collection dates for the SARS-CoV-2 data set used in this analysis. Data points corresponding to sequences collected in Ecuador are highlighted in red, and the linear regression trendline is shown in blue.

Table 1 provides details for each Ecuador transmission lineage (named sequentially according to the collection date of the earliest sequence in each lineage). We identify 82 (95% HPD: 81-84) SARS-CoV-2 introduction events from other countries into Ecuador through a robust counting approach (*43*); for comparison, this is slightly fewer than those reported for Brazil (~102 introductions) over a shorter time span during its first wave and with four times more sequences (*5*). This estimate assumes that transmission lineage groups D* and H* are comprised of 2 and 3 individual transmission lineages respectively (with an additional singleton inferred as part of H*; Fig 2A). The detection lag (defined as the number of days between the inferred transmission lineage TMRCA and its earliest sampled sequence) ranged between 1 and 140 days (Table 1), with a median of 16 days (IQR: 7-31 days; Fig 2B, Fig S10).

Table 1. Summary of transmission lineages identified in Ecuador.

Transmission lineage	Number of sequences (% of total sequences)	Earliest sample collection date	Latest sample collection date	Median TMRCA (95% HPD)	Detection lag***	Provinces where it has been sampled
A	3 (1.875%)	16/03/2020	23/03/2020	03/03/2020 (25/02/2020 - 09/03/2020)	13 days	Guayas
B	5 (3.125%)	30/03/2020	30/06/2020	16/03/2020 (02/03/2020 - 30/03/2020)	14 days	Imbabura, Los Rios, Pichincha
C	2 (1.25)	30/03/2020	30/03/2020	29/03/2020 (26/03/2020 - 30/03/2020)	1 days	Pichincha
D*	21 (13.125%)	07/04/2020	01/10/2020	02/03/2020** (19/02/2020 - 14/03/2020)	36 days**	Bolivar, Cañar, Carchi, Esmeraldas, Guayas, Imbabura, Loja, Los Rios, Morona Santiago, Pichincha, Tungurahua
E	3 (1.875%)	07/04/2020	30/06/2020	02/04/2020 (10/03/2020 - 07/04/2020)	5 days	Guayas, Manabi
F	6 (3.75%)	13/04/2020	10/10/2020	12/07/2020 (02/06/2020 - 28/07/2020)	23 days	Azuay, Bolivar, El Oro, Guayas, Loja
G	8 (5%)	17/04/2020	01/07/2020	01/04/2020 (25/02/2020 - 09/04/2020)	16 days	Chimborazo, Imbabura, Los Rios
H*	18 (11.25%)	17/04/2020	30/11/2020	21/02/2020** (11/02/2020 - 28/02/2020)	56 days**	Azuay, Cotopaxi, El Oro, Esmeraldas, Imbabura, Loja, Orellana, Pastaza, Santa Elena, Sucumbios, Zamora Chinchipe
I	3 (1.875%)	29/04/2020	08/07/2020	22/03/2020 (24/02/2020 - 08/04/2020)	38 days	Guayas
J	2 (1.25%)	26/05/2020	15/06/2020	22/05/2020 (03/05/2020 - 26/05/2020)	4 days	Napo
K	2 (1.25%)	29/05/2020	20/07/2020	25/05/2020 (13/05/2020 - 29/05/2020)	4 days	Guayas
L	5 (3.125%)	03/07/2020	19/08/2020	16/06/2020 (08/06/2020 - 02/07/2020)	17 days	Azuay, Loja, Napo, Orellana
M	3 (1.875%)	14/07/2020	13/08/2020	30/04/2020 (26/03/2020 - 18/06/2020)	75 days	Los Rios, Zamora Chinchipe
N	7 (4.375%)	14/07/2020	20/08/2020	20/06/2020 (19/05/2020 - 12/07/2020)	24 days	Azuay, Esmeraldas, Imbabura, Los Rios, Orellana, Pichincha, Zamora Chinchipe
O	3 (1.875%)	22/07/2020	07/09/2020	15/07/2020 (02/07/2020 - 22/07/2020)	7 days	Imbabura
P	3 (1.875%)	27/07/2020	04/11/2020	09/03/2020 (24/02/2020 - 08/04/2020)	140 days	Guayas
Q	2 (1.25%)	11/08/2020	26/11/2020	06/07/2020 (22/05/2020 - 24/07/2020)	36 days	Guayas, Pichincha
R	2 (1.25%)	12/08/2020	12/08/2020	09/08/2020 (20/07/2020 - 12/08/2020)	3 days	Imbabura
S	2 (1.25)	01/10/2020	05/10/2020	13/09/2020 (16/08/2020 - 01/10/2020)	18 days	Cotopaxi, Tungurahua
T	2 (1.25%)	05/11/2020	09/11/2020	27/10/2020 (05/10/2020 - 04/11/2020)	9 days	Guayas
U	2 (1.25%)	09/11/2020	09/11/2020	13/09/2020 (16/08/2020 - 01/10/2020)	57 days	Guayas
V	3 (1.875%)	28/11/2020	09/12/2020	21/11/2020 (01/11/2020 -27/11/2020)	7 days	Santo Domingo de los Tsachilas, Sucumbios

* Transmission complex, composed of two or more potential introductions of very genetically similar viruses.
** TMRCA of various transmission lineages and sequences outside of Ecuador; node not inferred in Ecuador.
*** Number of days between the inferred TMRCA and the earliest collection date in the transmission lineage

Size and persistence of transmission lineages

Initial molecular clock analyses showed our data set contains strong temporal signal overall, although many sequences from Ecuador showed lower than average genetic divergence from the root (Fig 2C). The inferred TMRCAs of Ecuadorian transmission

lineages ranged from February to November 2020 (Table 1); from this list, transmission lineages C and S are composed of pairs of sequences that share an epidemiological link.

The TMRCAs estimated for the two large transmission lineage groups D* and H* are the earliest in our data, however, these might not represent true lineage ancestors within Ecuador, because each group could represent more than one introduction from other countries. After excluding these larger groups, we still identified six transmission lineages for which the 95% HPDs of the TMRCA include the date of implementation of the national lockdown, March 16. Therefore, these transmission lineages likely correspond to introduction events that occurred before restrictions on incoming international flights were adopted in Ecuador. An additional four transmission lineages have TMRCA estimates between late March and November 2020. These may correspond to more recent introductions (following the progressive relaxation of the lockdown in Ecuador between May and September). The most likely exceptions to this are transmission lineages C (TMRCA: 2020.2412, 95% HPD: 2020.2328-2020.2446) and M (TMRCA: 2020.3278, 95% HPD: 2020.1864-2020.4187). Incomplete sampling of these lineages and detection lags could result in the date of introduction being substantially earlier than the date of the TMRCA (*46*), which would place the introduction date for these transmission lineages prior to the implementation of the lockdown.

Transmission lineages vary in size from sequence pairs (transmission lineages C, J, K, Q, R, S, T and U) to larger clusters of 16 to 21 sequences (transmission lineage group D*, depending on whether D* is considered as a single lineage or as multiple lineages).

The number of sequences in each transmission lineage is correlated with the number of days between the earliest and most recent sampling dates within the lineage (assuming that D* and H* are composed of multiple individual transmission lineages each). However, this result could be driven by the single largest transmission lineage in the data set. A similar pattern is observed when comparing the time between the inferred TMRCA and the most recent sequence of each transmission lineage (equivalent to the persistence time plus the detection lag; Fig S11). We also observed that lineages that were detected earlier tend to be larger (contain more sequences) and persist longer, but are not more geographically widespread (Fig 3A; Fig S12-S14). However, transmission lineages first detected or with a TMRCA between June and August are found on average in a greater number of provinces (from earliest detection of transmission lineages: *mean* = 3.27 provinces, *median* = 3 provinces; from TMRCAs of transmission lineages: *mean* = 3.7 provinces, *median* = 4 provinces) compared to transmission lineages first detected at any other time of the year (from earliest detection of transmission lineages: *mean* = 2.21 provinces, *median* = 1.5 provinces; from TMRCAs of transmission lineages: *mean* = 2 provinces, *median* = 1 province) (Fig 3A). Overall, transmission lineages with earlier TMRCAs and that were first detected earlier in the year persisted for longer timespans (i.e. there's a greater number of days between the first detection and the most recent detection of a transmission lineage; Fig S13). It should be noted however that these patterns also resemble the increased sampling that occurred over the months of July and August.

Geographical distribution of transmission lineages

Singletons and transmission lineages are found across multiple provinces and regions of Ecuador (Fig 3B-3C). Singletons represent an important proportion of the sequences in various provinces across central Ecuador ranging between 33.3% in Tungurahua and 52.3% in Guayas (Fig 3B), an observation that is particularly important for provinces with large numbers of sequences (Fig S14). On the other hand, different transmission lineages are found either in single provinces or across multiple regions (Fig 3C). The large transmission lineage groups D* and H* include sequences from provinces across three geographical regions each (the coastal region, the highlands and the Amazon region), and potentially show internal seeding events of the virus across provincial boundaries (Fig 2A, 3C). Even when accounting for the possibility that these lineage groups are comprised of multiple transmission lineages, sequences from different provinces and regions clustered together (Fig 2A).

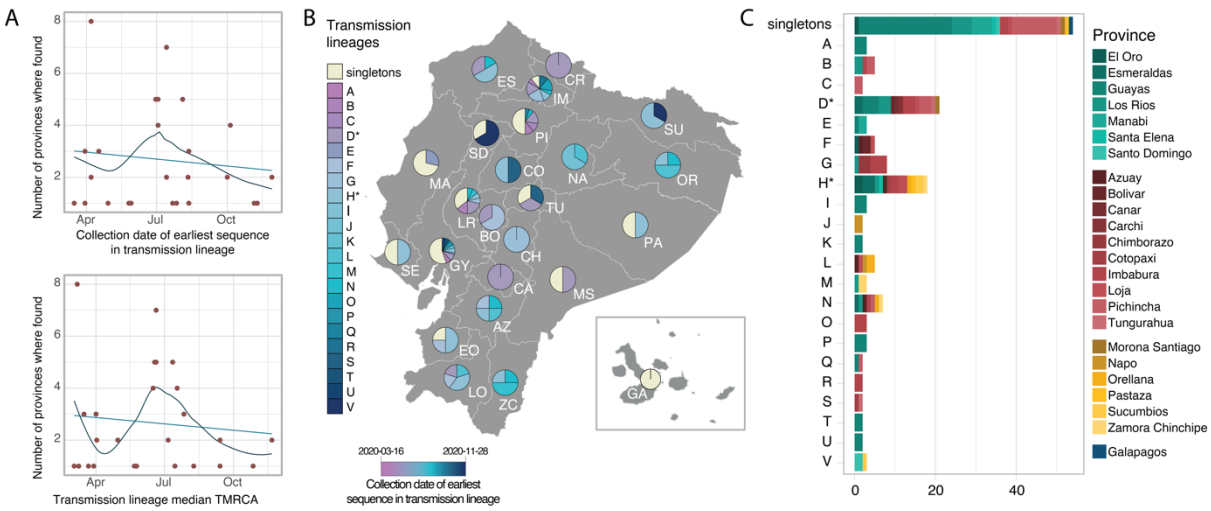


Figure 3. SARS-CoV-2 transmission lineages in Ecuador.

A. Summary of the geographic spread of transmission lineages in Ecuador, showing the number of provinces where each transmission lineage is found compared to the

collection date of the earliest sequence in each transmission lineage (upper panel) or the inferred median TMRCA for each transmission lineage (lower panel). The trend lines show a linear regression in light blue and a fitted local polynomial regression in dark blue. ***B.*** Contribution of individual transmission lineages and singleton sequences in each province. Transmission lineages (shades of blue) are ordered based on the earliest sample collection date in the group from earliest (darker) to more recent (lighter). ***C.*** Bar plot summarising the provinces where each transmission lineage was sampled over the study sampling period.

Consistent with the spatio-temporal sampling patterns (Fig 1C), the older transmission lineages (shown in darker blue in Fig 3B) were identified predominantly in provinces on the coast and highland regions, while younger transmission lineages (shown in lighter blue in Fig 3B) were identified in specific provinces in the north and more broadly in the south. The first epicentre of the COVID-19 epidemic in Ecuador, the province of Guayas, is represented by a high frequency of singleton lineages, with a high diversity of individual transmission lineages first identified at different times during 2020. A similar pattern is observed for the second epicentre of the epidemic, the province of Pichincha, but with fewer different transmission lineages and less representation of the youngest transmission lineages (Fig 3A). We note that these patterns could be affected by differences in the number of sequences available for each province (Fig 1A; Fig S15).

Discussion

The early weeks of the COVID-19 epidemic in Ecuador were characterised by a severe spike in the number of cases in city of Guayaquil, the largest in the country located in the province of Guayas, and by high attack rates (*i.e.* new cases in a population at risk divided by the size of that population at risk) across various coastal provinces

(*24*). The outbreak overwhelmed local healthcare systems, resulting in one of the highest excess death rates in the world during early 2020 (*23*). Information about the importation of SARS-CoV-2 into Ecuador and the domestic spread of the virus is needed to explain the drastic effects of the pandemic in the country during March and April 2020, and to explain the large difference in disease burden between Guayaquil and the capital, Quito.

An important determinant of the early dynamics of COVID-19 outbreaks has been human mobility and the number of introduction events of the virus into a new location with an immunologically naïve population, as was observed during the early stages of the emergence of SARS-CoV-2 in China (*47*). The large proportion of singletons observed in Guayas and various other coastal provinces, particularly given their tendency to occur early in the epidemic (Fig S2), could be suggestive of multiple independent introduction events with limited forward transmission. This would also explain why the earliest sequences in the local transmission lineages and the sequences assigned to the most common Pango lineage in Ecuador (B.1.1.74) were predominantly sampled in coastal provinces during the early weeks of the epidemic (Fig 2A, Fig S16).

Two additional factors support the hypothesis that Guayas played an important role in seeding of viral transmission to other regions in Ecuador: i) the city of Guayaquil hosts the second busiest international airport in the country and one of only two in the coastal region (the second international airport located in the province of Manabí hosts limited flights to a few international destinations; (*48*), and ii) the overall timing of the seeding events (which necessarily have to predate the inferred TMRCA of a lineage) corresponds to the school holiday period in the coastal region (February to April), when

international travel and large social gatherings are more likely to occur. The inferred TMRCAs, which can serve as an upper bound for the true introduction dates of the largest lineages (assuming these are better represented with the available sequences), show that these transmission lineages most likely arrived in Ecuador before the date when NPIs were implemented and before travel restrictions came into place.

The genetic diversity of SARS-CoV-2 can be quantified using the Pango dynamic nomenclature system. Pango lineages reflect the history of significant events in the epidemic and geographic spread of the virus (*49*) and can be used to explore likely source locations of virus importations, for example, the high representation of lineages descended from B.1.1 in the province of Guayas (Fig 1D). The emergence of B.1.1 in Europe and North America around February 2020 (*45*) might suggest these regions contributed to seeding the epidemic in this province of Ecuador (despite the limitations on identifying exact source locations due to poor surveillance in many countries). B.1.1.74, the most prevalent lineage in Ecuador, descended from B.1.1 and was more frequently sampled in Guayas during the early months of the epidemic (Fig S16). This further reveals that locations where B.1.1 was prevalent during the epidemic's first wave likely played a role as importation sources. However, the high proportion of singletons observed in Guayas also suggests that onward transmission of introduced virus was less common. Insights from regions sampled at very high intensities, such as the United Kingdom, show that the majority of introductions lead to small, transient, dead-end transmission lineages, whereas a smaller number of introductions lead to larger and longer-lasting transmission lineages (*15*). If this phenomenon is a general property of

the first wave of SARS-CoV-2 transmission, as appears to be the case given similar early observations in Brazil (*5*), Panama (*28*), Uruguay (*50*), Spain (*51*) and the Netherlands (*52*), we can expect that many of the introductions to Guayas led to few new cases, and that most of the ongoing transmission was derived from only a few introductions. This pattern of transmission heterogeneity would emerge due to the over-dispersion of SARS-CoV-2 transmission (*53, 54*), a phenomenon that could play an important role in the viral epidemic dynamics (*15, 55, 56*).

The larger transmission lineages identified here suggest that virus transmission was high between neighbouring and well-connected provinces. This might have been an important determinant of the transmission dynamics between the main cities in the country. In contrast to Guayaquil, Quito is the second-largest city in the country and presented a much less severe first epidemic wave, despite hosting the busiest international airport in the country. The city is located in the province of Pichincha, which exhibits a large proportion of singletons but fewer distinct global and transmission lineages overall. Importantly, the transmission lineages observed in Pichincha are mostly not shared with Guayas. This suggests that either independent international introductions or domestic seeding events likely drove the early epidemic in Pichincha. Our phylogenetic analyses suggest that some transmission lineages in Pichincha were introduced from other provinces (Fig 2A, e.g. the monophyletic Pichincha clade in lineage group D*; Fig S6), suggesting that domestic travel might have played an important role in the establishment of SARS-CoV-2 transmission in this region. It is also possible that a later burst of international introductions of new lineages occurred after

air travel and lockdown measures were lifted; however, the sampling dates of singleton genomes from Pichincha (between March and July) suggest that introductions into this province likely occurred before the lockdown and not during the relaxation of NPIs. The Pichincha singletons account for the earliest sequences of this kind in our data set, but could represent instances where limited or no additional spread occurred following their introduction. In fact, the overall heterogeneity across provinces for the number of circulating transmission lineages and singletons is noteworthy, and the possibility that other provinces account for additional ports of entry for the virus into the country could be further explored. For example, the province of Manabi (MA) also features an airport and an international maritime port and is represented predominantly by singletons (Fig 3B). Increased genomic coverage of such provinces would and facilitate a deper investigation into these patterns (at the current stage, the genomic representation is limited and likely underpowered for formal hypothesis testing).

Ultimately, more comprehensive analyses on the sources and drivers of transmission would require a deeper sampling of key locations where transmission was high, and the inclusion of complementary data sources such as real-time mobility and transportation data which could provide a better overview of the forces shaping the observed viral genetic diversity in Ecuador. Provinces such as Azuay, Guayas and Pichincha represent the main air travel entry points, but the role of land mobility across the northern border with Colombia and the southern border with Peru should also be considered to further understand the role of other Latin American countries in regional viral transmission.

Our analysis highlights some important patterns but is limited by various factors. Most notably, the number of genome sequences in our study, although large by historical standards, is small compared to the current trends for virus genome sequencing during the COVID-19 pandemic. The sample size restricts our ability to infer further details about local virus transmission dynamics from sequence data alone. The trajectories of individual lineages, from their international sources to their spread across the country and their subsequent local circulation, are best analysed from larger data sets, or in conjunction with additional data sources to manage and ameliorate the potential consequences of sampling biases (*10, 57*). Incorporating data from self-reported travel histories and human mobility can help to maximise the utility of smaller samples, collected in settings where the sequencing of large numbers of genomes lies beyond local technological or financial capacity, or where high sampling densities are unfeasible (e.g. in remote locations). Moreover, the broad range of lag times between the inferred TMRCAs and the earliest sampled sequence per transmission lineage suggest that a considerable number of transmission lineages are not detected immediately after being introduced to the country, reducing our capacity to identify a potential port of entry and delaying the possibility of responding rapidly to new seeding events. This can be particularly relevant for the identification of Variants of Concern (VOCs), where their potential for higher transmissibility (*58-60*) could require faster response times to identify the sources of importation and establish effective contact tracing to contain further spread.

The first year of the COVID-19 pandemic has shown how global connectivity plays a key role in the development of national epidemics caused by respiratory viruses, reminiscent of other pathogens such as Influenza viruses (*16*). Our results from Ecuador showcase the relevance of importations in establishing local viral circulation and the potential consequences of interprovincial mobility for highly connected locations. In particular, it shows that importations have been a common occurrence even after the implementation of lockdown measures and travel restrictions, and that seeding events across provinces can occur frequently. While air travel is limited between provinces, the connectivity provided by land travel can serve as a means for pathogen spread, highlighting the vulnerability of highly connected and remote locations alike. The notion that two large cities with busy international airports can manifest such different transmission dynamics and viral genetic diversity is also relevant, as it shows that a combination of multiple factors determines the outcome of an epidemic within a specific location. Ultimately, the type of interventions chosen to mitigate this high degree of connectivity, the necessity of an early implementation of these interventions and the adherence to these by the general population are paramount in determining their efficacy.

Data availability

The newly generated sequences have been publicly shared through the GISAID platform. The data sets (including Ministry of Health and National Institute of Statistics and Census data) and code used to generate the analyses presented in this study, as well as a list of accession numbers for the Ecuador SARS-CoV-2 genome sequences analysed

here are available through GitHub at https://github.com/BernardoGG/SARS-CoV-2_Genomic_lineages_Ecuador. We thank Andrés N. Robalino and Carlos Oporto for their contribution to the open licence Ecuacovid repository (https://github.com/andrab/ecuacovid) used to extract epidemiological data for this study. Trace plots representative of a Thorney BEAST run for selected parameters shown as Fig S16.

References

1. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, (2017).
2. J. L. Geoghegan *et al.*, Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat Commun* **11**, 6351 (2020).
3. C. Alteri *et al.*, Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat Commun* **12**, 434 (2021).
4. J. Lu *et al.*, Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003.e1009 (2020).
5. D. S. Candido *et al.*, Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255-1260 (2020).
6. M. Worobey *et al.*, The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564-570 (2020).
7. N. D. Grubaugh *et al.*, Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**, 10-19 (2019).
8. S. Duchene *et al.*, Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol* **6**, veaa061 (2020).
9. P. Lemey *et al.*, Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat Commun* **11**, 5110 (2020).
10. A. Kalkauskas *et al.*, Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Comput Biol* **17**, e1008561 (2021).
11. N. R. Faria *et al.*, Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406-410 (2017).
12. A. Rambaut, E. Holmes, The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Curr* **1**, RRN1003 (2009).
13. D. J. Park *et al.*, Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161**, 1516-1526 (2015).
14. L. W. Meredith *et al.*, Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* **20**, 1263-1271 (2020).
15. L. du Plessis *et al.*, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, (2021).
16. P. Lemey *et al.*, Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog* **10**, e1003932 (2014).

17. T. I. Vasylyeva, S. R. Friedman, D. Paraskevis, G. Magiorkinis, Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. *Infect Genet Evol* **46**, 248-255 (2016).
18. G. K. Moreno *et al.*, Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat Commun* **11**, 5558 (2020).
19. D. F. Gudbjartsson *et al.*, Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med* **382**, 2302-2315 (2020).
20. T. Sekizuka *et al.*, Haplotype networks of SARS-CoV-2 infections in the. *Proc Natl Acad Sci U S A* **117**, 20198-20201 (2020).
21. A. Popa *et al.*, Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci Transl Med* **12**, (2020).
22. N. D. Grubaugh *et al.*, Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic. *Cell* **178**, 1057-1071.e1011 (2019).
23. G. Long, Ecuador's virus-hit Guayaquil is grim warning for region. *Financial Times*. 2020.
24. E. Ortiz-Prado *et al.*, Epidemiological, socio-demographic and clinical features of the early phase of the COVID-19 epidemic in Ecuador. *PLoS Negl Trop Dis* **15**, e0008958 (2021).
25. H. Nishiura, G. Chowell, in *Mathematical and Statistical Estimation Approaches in Epidemiology,* G. Chowell, J. M. Hyman, L. M. A. Bettencourt, C. Castillo-Chavez, Eds. (Springer Netherlands, Dordrecht, 2009), pp. 103-121.
26. R. P. Fernández-Naranjo *et al.*, Statistical data driven approach of COVID-19 in Ecuador:. *Infect Dis Model* **6**, 232-243 (2021).
27. K. Laiton-Donato *et al.*, Genomic Epidemiology of Severe Acute Respiratory Syndrome Coronavirus 2, Colombia. *Emerg Infect Dis* **26**, 2854-2862 (2020).
28. D. Franco *et al.*, Early Transmission Dynamics, Spread, and Genomic Characterization of SARS-CoV-2 in Panama. *Emerg Infect Dis* **27**, 612-615 (2021).
29. S. Marquez *et al.*, Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19 in Ecuador. *medRxiv*, (2020).
30. D. Lopez-Alvarez, B. Parra, W. J. Cuellar, Genome Sequence of SARS-CoV-2 Isolate Cali-01, from Colombia, Obtained Using Oxford Nanopore MinION Sequencing. *Microbiol Resour Announc* **9**, (2020).
31. F. Wu *et al.*, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
32. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
33. B. Q. Minh *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
34. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).

35. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).
36. M. A. Suchard *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018).
37. M. A. Ferreira, M. A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains. **36**, 355-368 (2008).
38. M. S. Gill *et al.*, Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713-724 (2013).
39. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018).
40. X. Didelot, I. Siveroni, E. M. Volz, Additive Uncorrelated Relaxed Clock Models for the Dating of Genomic Epidemiology Phylogenies. *Mol Biol Evol* **38**, 307-317 (2021).
41. E. M. Volz, S. D. W. Frost, Scalable relaxed clock phylogenetic dating. *Virus Evolution* **3**, vex025 (2017).
42. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLoS Comput Biol* **5**, e1000520 (2009).
43. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol* **56**, 391-412 (2008).
44. C. J. Villabona-Arenas, W. P. Hanage, D. C. Tully, Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* **5**, 876-877 (2020).
45. A. Rambaut *et al.*, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020).
46. D. Duchêne, S. Duchêne, S. Y. Ho, Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Mol Ecol Resour* **15**, 785-794 (2015).
47. M. U. G. Kraemer *et al.*, The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493-497 (2020).
48. G. Hidalgo, J. León, R. Jácome, D. Vélez, "Anuario de Estadísticas de Transporte 2019," *Estadísticas de Transporte* (Instituto Nacional de Estadísticas y Censos, Quito, 2020).
49. N. M. Fountain-Jones *et al.*, Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol* **6**, veaa082 (2020).
50. V. Elizondo *et al.*, SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay. *Emerg Microbes Infect* **10**, 51-65 (2021).
51. F. Díez-Fuertes *et al.*, A Founder Effect Led Early SARS-CoV-2 Transmission in Spain. *J Virol* **95**, (2021).
52. B. B. Oude Munnink *et al.*, Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat Med* **26**, 1405-1410 (2020).

53. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355-359 (2005).
54. L. M. Li, N. C. Grassly, C. Fraser, Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol Biol Evol* **34**, 2982-2995 (2017).
55. L. Geidelberg *et al.*, Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evol* **7**, veaa102 (2021).
56. D. C. Adam *et al.*, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat Med* **26**, 1714-1719 (2020).
57. N. De Maio, C. H. Wu, K. M. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet* **11**, e1005421 (2015).
58. E. Volz *et al.*, Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, (2021).
59. H. Tegally *et al.*, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438-443 (2021).
60. N. R. Faria *et al.*, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, (2021).
61. A.F. Brito *et al.*, Global disparities in SARS-CoV-2 genomic surveillance. *MedRxiv Prepr. Serv. Heal. Sci.* 1–14 (2021).

4 Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America

Highlight: A successful recombinant lineage can be identified even when the parental lineages were *not systematically sampled* from the start

In this chapter, we use a combination of methods to investigate the emergence of a recombinant SARS-CoV-2 lineage which circulated in Mexico and the USA before the commencement of a systematic genomic surveillance programme in the former country. Genomic data, both from sequences and from re-occurring deletions show that lineage B.1.628 (now designated as XB) emerged from a recombination event. However, the detection of this lineage predates the detection of its inferred parental lineage, which highlights the effects of the relative frequency of a lineage in its detection through developing sampling programmes.

Abstract

While recombination is a feature of coronavirus evolution, previously detected recombinant lineages of SARS-CoV-2 have shown limited circulation thus far. Here, we present a detailed phylogenetic analysis of four SARS-CoV-2 lineages to investigate the possibility of virus recombination among them. Our analyses reveal well-supported

phylogenetic differences between the Orf1ab region and the rest of the genome (S protein and remaining reading frames). By accounting for several deletions in the NSP6, Orf3a and S, we conclude that the B.1.628 major cluster, now designated as lineage XB, originated from a recombination event between a B.1.631 virus and a lineage B.1.634 virus. This scenario is supported by the spatiotemporal distribution of these lineages across the USA and Mexico during 2021, suggesting this region is where the recombination event took place. This event raises important questions regarding the role and potential effects of recombination on the evolution of SARS-CoV-2 during the ongoing COVID-19 pandemic.

Introduction

Virus recombination, the process by which genetic material from two genetically distinct parental lineages is combined into a viable descendant virus genome, is a common feature of sarbecovirus evolution (Boni et al, 2020). Genomic analyses suggest that recombination events among coronaviruses circulating in non-human species occurred during the evolutionary history of SARS-CoV-2 (Zhu et al, 2020; Li et al, 2020; Lytras et al, 2021). Signals of ongoing recombination among SARS-CoV-2 genomes assessed under a statistical framework during the COVID-19 pandemic have been reported (VanInsberghe et al., 2021). Notably, SARS-CoV-2 genomes that are clearly recombinant have been observed at low frequencies in the UK, some of which showed evidence of forward transmission (Jackson et al, 2021). One of these UK recombinants was designated as lineage XA, the first recombinant lineage in the Pango nomenclature system (O'Toole et al., 2021; Rambaut et al., 2020). More recently,

potential recombinants between two variants of concern (VOCs), Alpha and Delta, have been detected in a small cluster in Japan (Sekizuka et al., 2021). Although few clearly recombinant SARS-CoV-2 lineages have been reported so far, our ability to detect them is likely to increase as time progresses, due to the continued genetic divergence of SARS-CoV-2 and increased co-circulation of divergent lineages.

Our understanding of the effects of genomic recombination on SARS-CoV-2 fitness and transmission dynamics are still limited, but genetic exchange has been previously associated with evolutionary adaptation in viruses under experimental conditions (e.g. Poliovirus; Xiao et al., 2016), in individual hosts (e.g. HIV; Song et al., 2018) and in nature (e.g. human influenza viruses; Petrova & Russell, 2018). Interestingly, a recombination event is associated with the emergence of a MERS-CoV lineage that became dominant in camels in the Middle East between 2014 and 2015 (Sabir et al., 2016). However, the question regarding the potential for recombination to contribute to SARS-CoV-2 evolution and adaptation remains. The emergence of highly divergent variants also raises questions regarding the role of recombination in the occurrence of large sequence shifts. While there is currently no evidence suggesting that recombination played a role in the origins of the recently designated VOC Omicron (Pango lineage B.1.1.529; Callaway, 2021; Technical Advisory Group on SARS-CoV-2 Virus Evolution, 2021), the accumulation of substantial numbers of mutations as observed in this variant could be produced by recombination mechanisms (Awadalla, 2003). It is therefore key to analyse the extent to which viral recombination drives adaptation in SARS-CoV-2.

For virus recombination to occur, the parental lineages need to co-circulate in the same location in order to allow specific individuals to become co-infected. This scenario provides the circumstances during which chimeric genotypes can emerge, usually through molecular processes such as template switching, homologous recombination or reassortment (the latter occurring in viruses with segmented genomes; Simon-Loriere & Holmes, 2011). Coronaviruses naturally produce a variety of recombination products during natural infection, including recombinant genomes, a process mediated by the proofreading exoribonuclease (Gribble et al., 2021).

Mosaic genomes resulting from recombination can be detected through changes in sequence similarities among different regions of the virus genome relative to parental lineages. Identifying recombination between recently diverged lineages is difficult, because sequence similarities are high, and it is hard to distinguish homoplasic changes from those that are identical by descent due to inheritance from a recent shared ancestor (synapomorphic changes). In such instances, other mutations such as insertions and deletions can prove informative; specifically, deletions are highly unlikely to revert during the evolution of a single lineage. Phylogenetic methods can also provide a tractable approach to test hypotheses regarding virus recombination, as they can be used to reconstruct the separate evolutionary histories of subgenomic regions (Simon-Loriere and Holmes, 2011). While genome regions that share the same ancestry can be easily established for segmented viruses (e.g., Orthomyxoviruses, such as influenza viruses [Holmes et al, 2005]), the exact start and endpoints of recombinant genome regions must be inferred statistically for non-segmented viruses (Pérez-Losada et al, 2015).

Furthermore, estimating the timing and location of recombination events can be limited by uncertainty in estimates of phylogenetic node ages, although such uncertainty can be reduced by using methods that combine evolutionary information across different genome regions (e.g., Raghwani et al, 2012).

As SARS-CoV-2 circulates around the world, new lineages emerge and are tracked using the Pango dynamic hierarchical nomenclature system (Rambaut et al, 2020). During late 2020 and early 2021, a series of lineages descending from B.1 were first detected in North and Central America. Specifically, lineages B.1.627, B.1.628, B.1.631 and B.1.634 were detected by the national genomic surveillance programs in the United States of America, Mexico and other countries in the Americas and their genomes shared publicly on the GISAID database (Shu & McCauley, 2017). Genomic similarities among these and other lineages circulating in the region and elsewhere prompted the suggestion that recombination had occurred during their emergence and spread (this hypothesis was initially proposed and discussed on the Pango GitHub website: https://github.com/cov-lineages/pango-designation/issues/189).

Here, we present an analysis of the spread and evolution of these four lineages and investigate the possibility that one or more recombination events contributed to their evolution. Our results provide evidence supporting a recombinant origin for lineage B.1.628 and its designation as a distinct recombinant lineage that presented forward transmission, circulating in multiple countries.

Methods

Genomic data, metadata and sequence alignment

We downloaded complete genome sequences from GISAID (Shu and McCauley, 2017) from individual Pango lineages detected in countries from Central America and Mexico (as of August 12 2021) and generated a consensus list of mutations for each lineage. This was done by extracting individual mutations per genome in relation to the SARS-CoV-2 reference genome Wuhan-Hu-1 (Wu et al., 2020) using the Augur pipeline (Hadfield et al., 2018) and identifying the ones that were found in >85% of the sequences assigned to said lineage. Each entry in this consensus mutation list identifies the locus, position (in reference to Wuhan-Hu-1) and type of nucleotide change. From these consensus mutation lists for each lineage, we generated a pairwise matrix of the number of shared mutations between individual lineages (i.e., absent in Wuhan-Hu-1 and shared by individual pairs of lineages). The results from these preliminary analyses have been presented as a Twitter thread at https://twitter.com/shay_fleishon/status/1425775733167820814 and as a Github Issue at https://github.com/cov-lineages/pango-designation/issues/189.

Lineages B.1.627, B.1.628, B.1.631 and B.1.634 displayed unusually high numbers of shared mutations; we therefore retrieved all complete SARS-CoV-2 genome sequences assigned to these four Pango lineages as of August 30 2021 from GISAID (Shu and McCauley, 2017). Accompanying sequence metadata, including sampling locations (at different geographic resolutions) and dates of sample collection and submission were also retrieved. This complete data set included 1950 sequences that

were subsequently filtered to exclude all sequences for which >10% of sites were ambiguous (i.e., had nucleotide states N or X). The final data set, comprising 1055 genome sequences, was used for all phylogenetic analyses. The original complete data set (n=1950) was used in part to explore the spatio-temporal distribution of the four Pango lineages under investigation.

After adding the reference SARS-CoV-2 genome Hu-1 (GenBank accession MN908947.3; Wu et al 2020) to the filtered data set, a single sequence alignment was constructed using MAFFT v7.487 (Katoh & Standley, 2013). The resulting alignment was inspected visually to identify deletions >1nt in length and which were shared by all or most of the sequences assigned to one or more of the lineages under investigation; these deletions were masked from the alignment and encoded as discrete sequence traits, which were later mapped onto estimated phylogenetic trees (see *Results*).

Confirmation of Pango lineage assignment and whole genome phylogenetic analysis

The Pango lineage assignment of the sequences in our final data set was determined using Pangolin v.3.1.11 (Rambaut et al., 2020); all lineages originally assigned on GISAID were confirmed. To further explore the phylogenetic structure of our data, we constructed a maximum likelihood (ML) phylogenetic tree using IQtree (Minh et al., 2020) under a GTR+Γ substitution model. We estimated node support values using the Shimodaira-Hasegawa (SH) approximate Likelihood Ratio test (SH-aLRT; Guindon et al., 2010), with 1000 replicates, and 1000 bootstrap replicates.

Genome-wide divergence of Pango lineages and recombination breakpoint inference

Pairwise genetic distances across the length of the genome of basal sequences for each of the lineages under investigation and the reference Wuhan-Hu-1 genome were estimated using custom scripts (available at https://github.com/BernardoGG/XB_lineage_investigation), based on the ape package in R (Paradis and Schliep, 2019). One basal sequence was selected per lineage (to represent all lineage-defining mutations), and in cases where multiple important clades were identified within single a Pango lineage one basal sequence from each of these clades was included. Raw genome-wide distances were estimated for 500-nucleotide segments, and a sliding window approach was used to estimate these distances across segments that overlapped every 20 nucleotides.

Recombination tests are computationally demanding for large data sets. To address this, we further subsampled the alignment to include one sequence per country per Pango lineage per day (script available at https://github.com/BernardoGG/XB_lineage_investigation), resulting in a downsized set of 716 whole genome sequences. To evaluate recombination patterns in our data, we further reduced the downsized data set by randomly sampling 200 sequences from the four lineages under investigation. This reduced data set was then analysed using GARD (Kosakovsky Pond et al, 2006a), a tool that uses a genetic algorithm to search for one or more putative breakpoints in a multiple sequence alignment; the best supported number of non-recombinant fragments in the alignment is then evaluated by comparing the

Akaike Information Criterion (AIC_c) of the proposed models versus a null model (i.e., no recombination points, such that a single tree topology best explains the sequence alignment). The phylogenetic topological incongruencies of potential non-recombinant fragments can be evaluated to identify lineages involved in the recombination event (Kosakovsky Pond et al, 2006b). We then compared the results obtained from GARD with those obtained using other recombination detection methods with a different approach. For this, we ran RDP4 (Martin and Rybicki, 2000; Martin et al., 2015), a computer program that sequentially tests every combination of three sequences in an alignment to find evidence that one of the three sequences is a recombinant and the other two are parental. We analysed the aforementioned 716-sequence alignment using the methods implemented in RDP4: Bootscan, MaxChi, Chimaera, 3Seq, GENECONV, SiScan and RDP (Martin et al., 2015). Statistical tests are unique to each recombination detection method in RDP4, and approximate *p*-values are reported where evidence for recombination was identified.

Phylogenetic analyses of inferred non-recombinant genome segments

Given the possibility that a single phylogeny of the complete genome alignment does not explain the evolutionary history of our sequences, we partitioned the alignment using the inferred breakpoints from GARD (derived from the analysis of the 200-sequence data set). ML trees for each genome partition were then estimated with IQtree as described previously.

We identified four key deletions amongst our sequences under investigation: two deletions in the NSP6 gene (Orf1ab) and two deletions in Orf3a (see *Results*). NSP6 deletions occur at two adjacent locations (here called ΔR1 and ΔR2) and don't result in changes in the reading frame. Orf3a deletions occur on a single locus and take the form of either a single-nucleotide frameshift deletion (ΔFS2) or a 4-nucleotide frameshift deletion (ΔFS4). These deletions were coded as discrete characters, assigned to individual tree tips and reconstructed at internal nodes using a parsimony criterion, thereby visualising the history of their occurrence across the phylogenies; these patterns have to reconcile with the proposed recombination events. Our rationale is that the individual evolutionary histories of each of the putative non-recombinant fragments should also parsimoniously explain the occurrence of deletions observed in NSP6, S and ORF3a, under the assumption that deletions do not revert once they occur in a lineage. Our rationale also draws from the premise that these deletions are more likely to descend from single occurrences within the evolution of each lineage but are not restricted to have occurred just once across the whole phylogeny. Specifically, some of these deletions have been observed previously in other lineages and variants of concern, including the NSP6 deletions observed in B.1.1.7, P.1 and B.1.351 (Meng et al., 2021). The genomic position where a given deletion occurs (relative to the inferred recombination breakpoints) was used to determine which genome partition most likely represents the true evolutionary history of that deletion. Loci where the deletion was flanked by ambiguities were differentially labelled with an asterisk (*).

Exploring the phylogenetic discrepancies of the lineages under investigation relative to other B.1 lineages

Our phylogenetic analyses showed that lineages B.1.628 and B.1.631 are split into two groups each. B.1.628 contains a cluster of sequences that fall near the root of the phylogenies and are henceforth identified as *B.1.628 minor*, and a large more derived monophyletic clade henceforth identified as *B.1.628 major*. B.1.631 is split into a small cluster of sequences that consistently cluster near the tree backbone in both genome segments and is henceforth identified as *B.1.631 minor,* and a larger monophyletic clade here called *B.1.631 major*. To explore the topological discrepancies between the lineages under investigation in the absence of B.1.631 minor (see *Results* for an explanation of why this cluster was excluded), we randomly sampled five sequences from each of the lineages under investigation (B.1.627, B.1.628 major, B.1.628 minor, B.1.631 major and B.1.634), with five random sequences from the B.1.1.7 lineage (i.e. VOC Alpha) and five random sequences from the B.1.351 lineage (i.e. VOC Beta); this was performed in order to explore the congruency of the diversification patterns of the lineages under investigation in context of the B.1 lineage. Five additional random sequences from lineage A.2.5 were also included to represent the Pango A lineage and to provide an outgroup for tree rooting. All sequences from the B.1.1.7, B.1.351 and A.2.5 lineages were obtained from GISAID; they had sampling dates that spanned the time when these lineages were observed to circulate and were predominantly from locations in North and Central America. We performed GARD analyses and constructed ML phylogenetic trees from this data set as previously described. We estimated node

support for the phylogenetic analyses using 1000 bootstrap replicates, and nodes with >50% node support were collapsed into polytomies.

Finally, we used the snipit software (https://github.com/aineniamh/snipit) to explore the distribution of single nucleotide polymorphisms (SNPs) across the genome of potential recombinant and parental lineages. SNPs were identified and visualised in reference to the Wuhan Hu-1 genome sequence.

Results

Distribution of lineages B.1.627, B.1.628, B.1.631 and B.1.634

We noted the global spatiotemporal distribution of a total of 1950 sequences for lineages B.1.627 (n = 252), B.1.628 (n = 1391), B.1.631 (n = 181) and B.1.634 (n = 126). Sequences from the four lineages were collected between July 8 2020 and August 18 2021 (Table 1), however the majority were sampled in 2021 (Fig 1A). All four lineages were predominantly sampled in North America (89.5% of sequences), either in the United States of America (USA) or Mexico. B.1.627 and B.1.631 were mostly sampled in the USA whereas B.1.634 was commonly found in Mexico (Fig 1B). Lineage B.1.628 is the most geographically widespread lineage in our data set: it has been identified in 41 different USA states and in 31 Mexico states (all other lineages were identified in 10-21 USA states and in 15-17 Mexico states; Fig 1C). B.1.628 is also the most widely sampled through time, with 406 days between the earliest and most recent sample collection date (compared to B.1.627 = 212 days, B.1.631 = 232 days, B.1.634

= 160 days). B.1.628 was sampled in the USA during the entirety of this sampling period, while only sampled for a period of 185 days in Mexico.

Table 1. Summary of sequences from four SARS-CoV-2 lineages associated with potential recombination (percentages shown in reference to the complete number of sequences in our data set).

	B.1.627		**B.1.628**		**B.1.631**		**B.1.634**	
	N (%)	Date range	N (%)	Date range	N (%)	Date range	N (%)	Date range
Africa	0 (0%)	---	0 (0%)	---	20 (1.0%)	2021-05-21 to 2021-07-12	2 (0.1%)	2021-05-29
Asia	4 (0.1%)	2020-12-21 to 2021-02-28	16 (0.8%)	2020-11-15 to 2021-05-23	1 (0.05%)	2021-07-07	0 (0%)	---
Caribbean Islands	0 (0%)	---	3 (0.15%)	2021-06-02 to 2021-07-03	0 (0%)	---	1 (0.05%)	2021-03-11
Central America	46 (2.4%)	2021-01-04 to 2021-07-08	58 (3.0%)	2021-01-26 to 2021-07-26	11 (0.6%)	2021-04-23 to 2021-04-30	1 (0.05%)	2021-06-15
Europe	0 (0%)	---	24 (1.2%)	2021-01-04 to 2021-08-05	14 (0.7%)	2021-02-18 to 2021-08-05	0 (0%)	---
North America	202 (10.4%)	2021-01-16 to 2021-07-21	1287 (66.0%)	2020-07-08 to 2021-08-18	134 (6.9%)	2020-12-16 to 2021-08-04	122 (6.3%)	2021-08-09
Oceania	0 (0%)	---	1 (0.05%)	2021-05-24	0 (0%)	---	0 (0%)	---
South America	0 (0%)	---	2 (0.1%)	2021-04-29 to 2021-08-03	1 (0.05%)	2021-05-28	0 (0%)	---
TOTAL	12.92%	2020-12-21 to 2021-07-21	71.33%	2020-07-08 to 2021-08-18	9.28%	2020-12-16 to 2021-08-05	6.46%	2021-03-02 to 2021-08-09

A maximum likelihood (ML) tree inferred from these genomes and rooted on the reference genome Wuhan Hu-1 shows that all four lineages form monophyletic clusters as expected (Fig. S1). Two exceptions are noted for lineages B.1.628 and B.1.631. For the former, a group of sequences close to the root of the tree are designated as lineage B.1.628 and this group is distinct from the main B.1.628 clade. In the latter case, sequences from lineage B.1.631 are split into two paraphyletic clusters by B.1.627 (Fig

S1 inset). For reference purposes within this work, we henceforth refer to the larger monophyletic B.1.628 and B.1.631 clades as *B.1.628 major* and *B.1.631 major*, and identify the sequences clustering at the base of the phylogeny that were assigned to these lineages as *B.1.628 minor* and *B.1.631 minor*. Some of the nodes that define important splits in the tree show moderate support; for example, the node that groups most B.1.631 genomes with other B.1.627 genomes (to the exclusion of the outlying B.1.631 genomes) is well supported (SH-aLRT = 98.6; Fig S1). The presence of these phylogenetic clusters that don't match the Pango lineage definitions (Rambaut et al., 2020) warrants further investigation, with recombination as a possible explanatory factor.

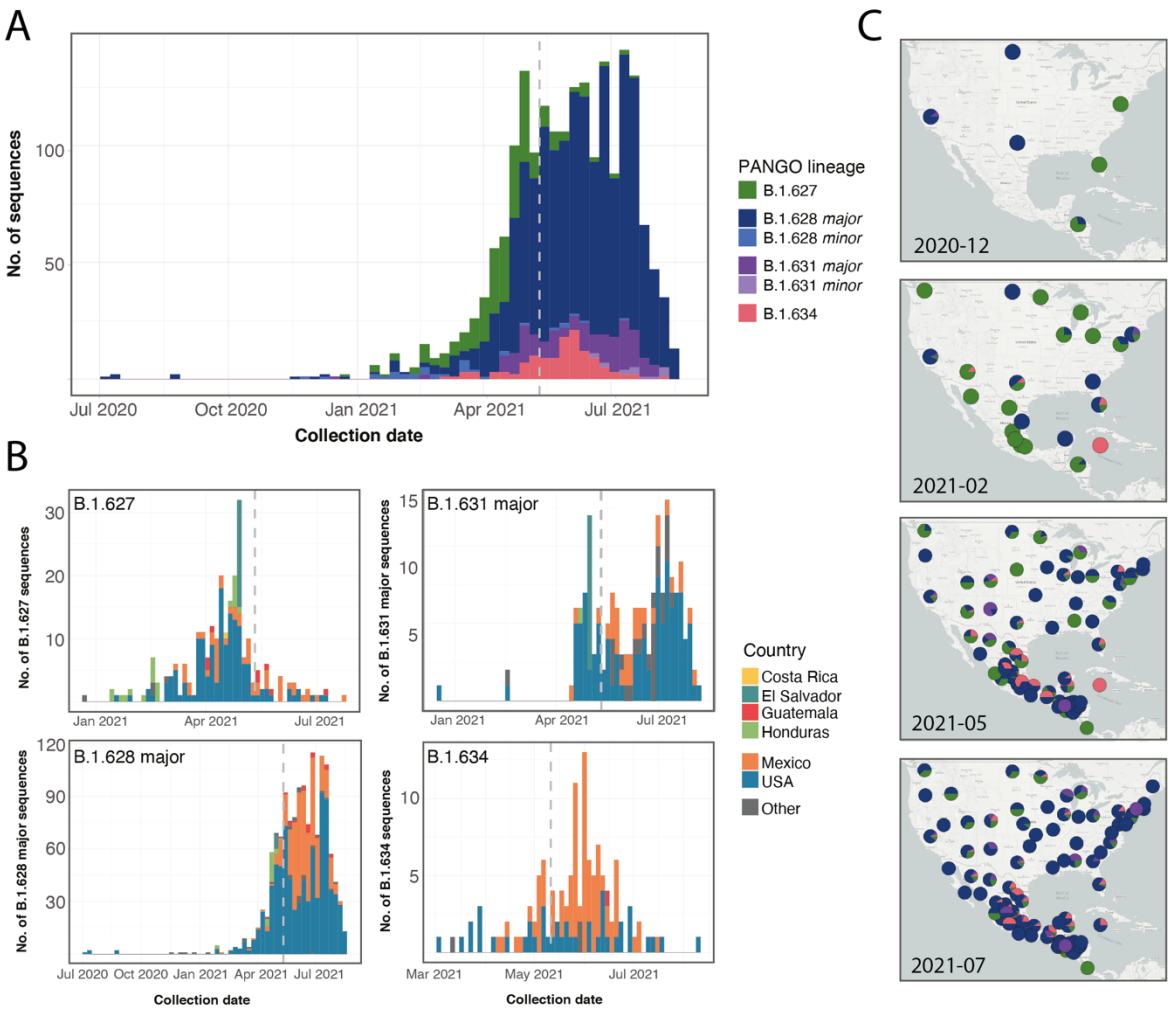


Figure 1. Spatiotemporal distribution of lineages B.1.627, B.1.628 (major and minor), B.1.631 (major and minor) and B.1.634. (A) Number of sequences on GISAID (per two-week period, as of 2021-08-30) for each lineage publicly available in GISAID, plotted using the associated sample collection date for each sequence. **(B)** Weekly number of sequences for each lineage coloured by country from where samples were collected. Countries outside of North and Central America make up <5% of sequences and are therefore grouped in the "Other" category. The dotted vertical lines show the starting of the systematic genome sampling and sequencing programme according to the national SARS-CoV-2 genome surveillance program for Mexico (May 11, 2021). B.1.628 minor and B.1.631 minor lineages are not shown. **(C)** Mapping of the geographic spread of the four lineages in North and Central America (a region where >95% of the sequences were identified) at four representative dates over their complete sampling date range.

Recombination analyses and breakpoint inference

Our results indicate that recombination is likely to have occurred in our data set. The GARD analysis suggests that a single breakpoint in the alignment can explain the data (Fig 2A), with high support for a model incorporating this recombination event ($\Delta c\text{-}AIC_{null\ model} = 202.176$). The breakpoint is inferred by GARD to have occurred around position 21308 (a TTT codon), corresponding to the signal peptide region at the N-terminus of the spike protein (18 nucleotides downstream of the canonical sarbecovirus transcription regulatory sequence AACGAAC; Yang et al., 2021). However, some variation in the results is observed when using different subsampling approaches; an analysis that excludes the B.1.634 lineage indicates a recombination breakpoint is inferred position 22775-22778 (at a GAT codon) in the Spike protein reading frame ($\Delta c\text{-}AIC_{null\ model} = 562.098$), corresponding to amino acid residue 390D located in the core region of the receptor binding domain (RBD) adjacent to beta sheet 3 (β3; Lan et al., 2020). Moreover, the RDP4 results for the B.1.627 and B.1.628 sequences revealed a recombination breakpoint inferred at position ~19408/19411 ($p_{MaxChi} < 0.0001$, $p_{3Seq} < 0.0001$), corresponding to the N7-MTase domain of nsp14, at the end of the Orf1ab. In any scenario, the receptor binding motif (RBM) which includes the main ACE2 receptor contact points would have been inherited from the same parental lineage (B.1.628).

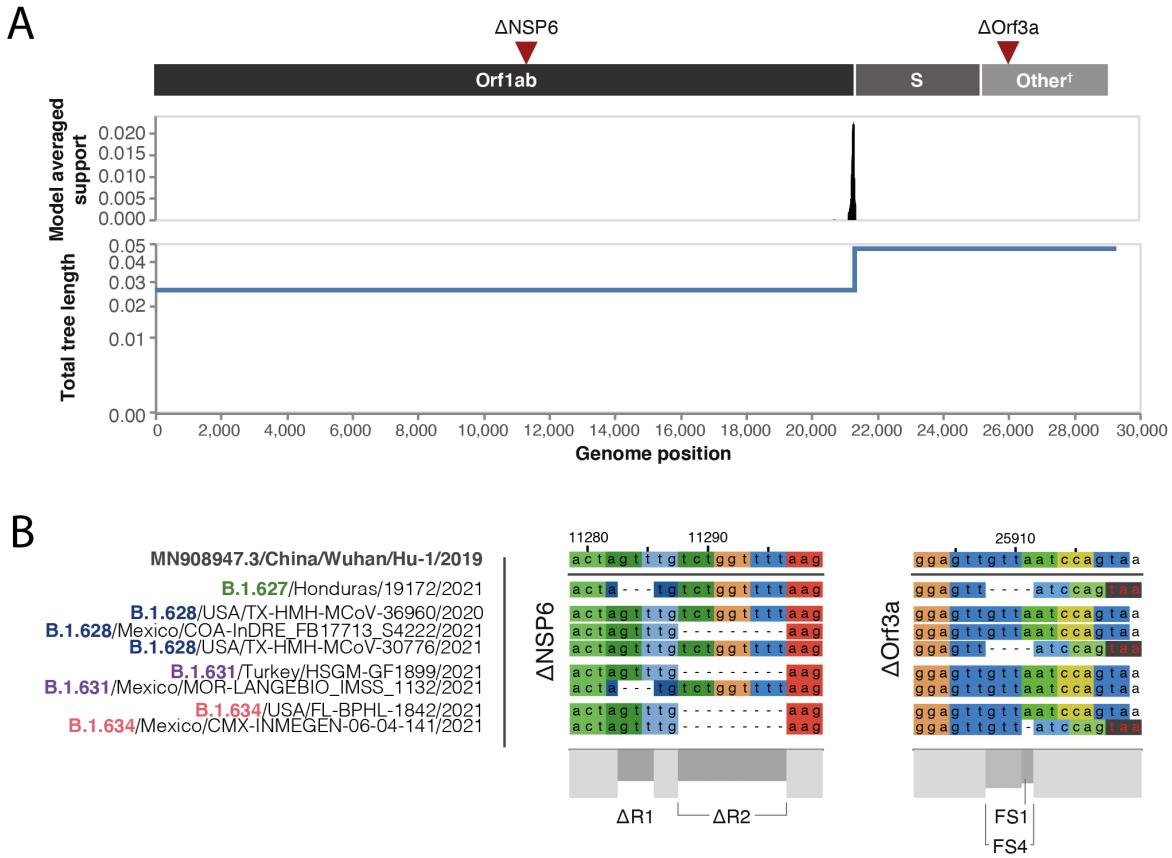


Figure 2. Recombination breakpoint analysis and deletions occurring in the four lineages. (A) Recombination breakpoint analysis results performed on GARD. The upper panel shows high support for a model where two genome segments split at position 22775 best explain the evolutionary history of the sequences from the lineages under investigation. The lower panel shows the different total tree lengths for the phylogenies of each of the inferred genome segments. The genomic location of deletions under investigation (ΔNSP6 and ΔOrf3a) are shown for reference on top. **(B)** Deletions in the NSP6 gene (Orf1ab) and Orf3a, illustrated on a representative selection of sequences that includes the B.1.627, B.1.628, B.1.631 and B.1.634 lineages. NSP6 deletions (ΔR1 and ΔR2) and Orf3a deletions (a single-nucleotide frameshift deletion ΔFS2 or a 4-nucleotide frameshift deletion (ΔFS4) are shown, with the early TAA stop codon produced by the Orf3a deletions shown in red letters on a black background.

The recombination analysis outcome also results in the placement of the NSP6 deletions on one side of the breakpoint, and the Orf3a deletions on the other side of the

breakpoint (Fig 2A, 2B). The NSP6 region contains two non-overlapping deletions: a 3-nucleotide deletion (ΔR1) and a 6-nucleotide deletion (ΔR2) which are 2 nucleotides apart, but do not derivate in frame shifts downstream (Fig 2B). The Orf3a displays two overlapping frameshift deletions: a single-nucleotide deletion (FS1) and a 4-nucleotide deletion (FS4) which overlap on the fourth position of FS4. Both deletions lead to the same early TAA/UAA termination codon, six nucleotides downstream of the FS1/FS4 locus (Fig 2B).

Phylogenetic discrepancies between non-recombinant genome segments

Given the inferred recombination breakpoint close to the start of the S reading frame (Fig 2A), we estimated separate phylogenetic trees for Orf1ab and for the remainder of the genome (including S and the remaining structural and non-structural genes, henceforth referred to as the S-3' region). The phylogenies show topological discrepancies which coincide mostly with individual Pango lineages (Fig 3). Both phylogenies (rooted on the Hu-1 reference genome) show a poorly resolved early split, with Orf1ab showing a bifurcation into two monophyletic groups containing B.1.628 minor and B.1.631 minor basal sequences. On the other hand, the phylogeny for the S-3' region places the B.1.628 minor sequences as a paraphyletic group from which B.1.634 descends (SH-aLRT = 91.9), whilst B.1.631 minor is a predecessor of the B.1.627, the B.1.628 major and the B.1.631 major lineages (SH-aLRT = 79.7). On the Orf1ab phylogeny, lineages B.1.627 and B.1.631 major share a common ancestor (SH-aLRT = 86.5), while the S-3' tree shows them as being paraphyletic. Lineage B.1.634 is

consistently inferred as monophyletic in both trees; in the Orf1ab tree it descends from B.1.631 minor, and in the S-3' tree from B.1.628 minor. The nodes defining lineages are generally well supported (SH-aLRT > 70.0), except for the basal nodes for the early bifurcation in both trees; statistical support within each lineage showed a combination of unsupported short branches (SH-aLRT = 0.0) and nodes with high support values (SH-aLRT > 75.0).

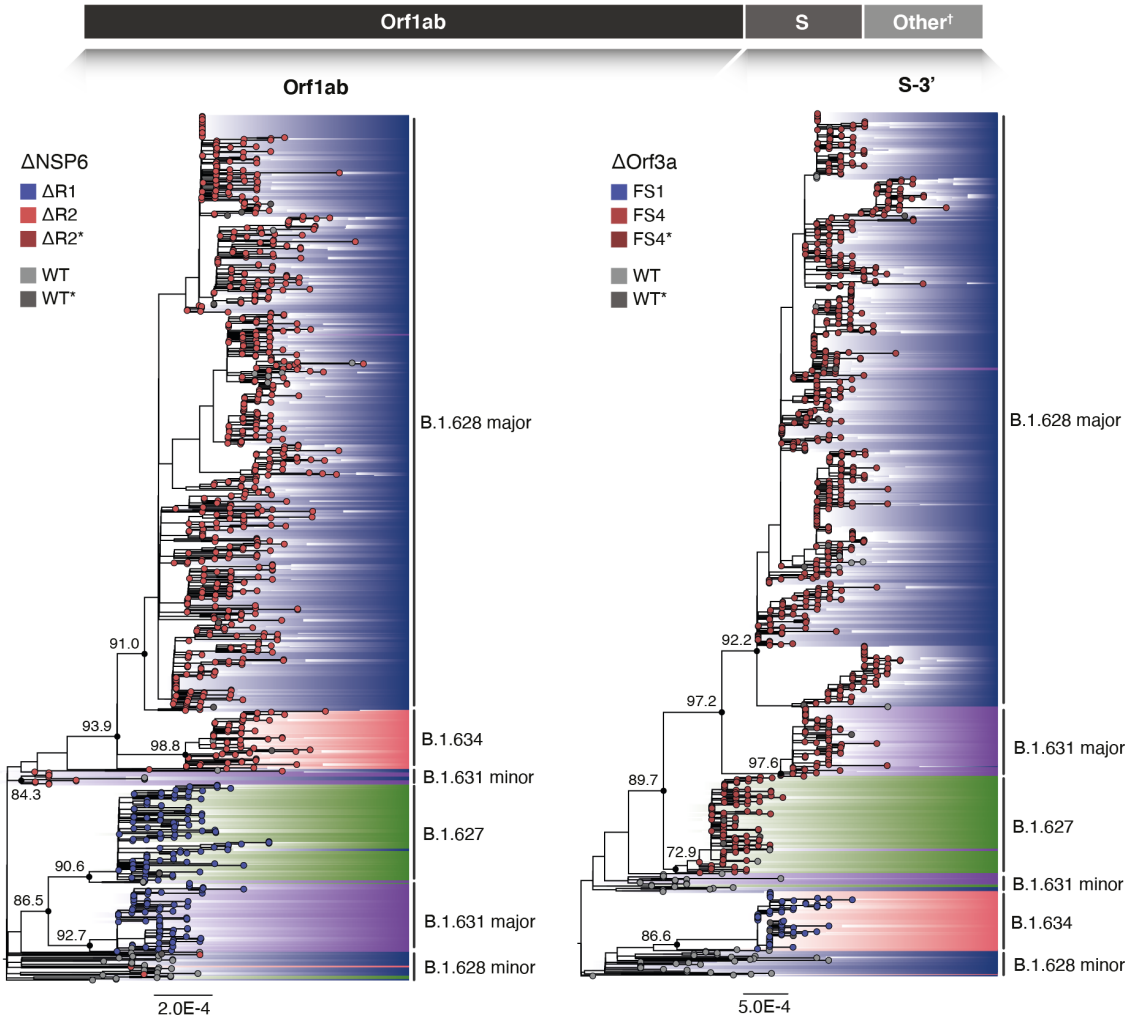


Figure 3. Maximum likelihood phylogenies of two segments of the SARS-CoV-2 genome for the four lineages. Individual phylogenies were constructed for the 5' and

3' end of the genome, split at the inferred breakpoint, resulting in trees that represent the evolutionary histories of Orf1ab (left panel) and the S gene plus the remaining structural and non-structural genes† (here referred to as S-3', right panel). The individually designated Pango lineage for each sequence is highlighted (the consensus identifier for each clade is also shown), and SH-aLRT node support is shown for key lineage defining nodes on both phylogenies. Deletions occurring on each genome segment (ΔNSP6 on Orf1ab and ΔOrf3a on S-3') are mapped onto the tips of the trees. †Other genes: S, E, M, N, Orf3a, Orf3b, Orf6, Orf7a, Orf7b, Orf8, Orf9b, Orf9c and Orf10.

Genome-wide divergence across genomes

To further explore the genetic divergence of these lineages and the sequences placed near the root of the trees (specifically, B.1.631 minor and B.1.628 minor), we estimated the pairwise genetic distances across the genomes of representative sequences (basal to the main clades) in reference to the Hu-1 reference genome (Fig S2). While mutations have accumulated in all lineages, the Orf1ab region of B.1.631 minor shows the lowest divergence from Hu-1; all clades display peak genetic divergence between positions ~21000 and ~23000 (an expected pattern as this region involves the highly variable S gene), with the exception of B.1.628 major which diverges from Hu-1 homogeneously across its genome.

Emergence of lineages from a B.1 background and recombination history

Given the genome-wide divergence observed for B.1.631 minor (particularly the unusually high similarities to Hu-1 in the Orf1ab region) and its limited spatiotemporal distribution (i.e. all sequences are from Turkey as opposed to the majority of the sequences that were observed in North and Central America), we excluded this group from further analyses. In particular, the fact that it was not observed anywhere in the

Americas suggests that it would not have circulated in the same geographical region, a necessary condition for recombination to occur. In the absence of this cluster, we explored the evolutionary history of the remaining lineages in relation to other lineages that descend from B.1. A phylogenetic analysis including a sample from each lineage under investigation, B.1.1.7 (VOC Alpha) and B.1.351 (VOC Beta) consistently shows that B.1.627, B.1.628 major, B.1.631 major, and B.1.634 group into well-supported monophyletic groups (bootstrap support >93%) for both the Orf1ab (Fig S3) and the S-3' (Fig S4) genome segments, similar to the well-established Alpha and Beta VOCs. The B.1.628 minor sequences emerge from the polytomy that makes up the B.1 backbone and do not group with the B.1.628 major cluster, suggesting the former do not belong to the B.1.628 Pango lineage. The Orf1ab phylogeny shows that B.1.627 and B.1.631 major share a common ancestor (bootstrap support = 62%), similar to B.1.628 major and B.1.634 (bootstrap support = 100%; Fig S3); this pattern is also observed in the full phylogeny (Fig 3). The S-3' tree shows that B.1.628 major and B.1.631 major share a common ancestor (bootstrap support = 95%) which in turn descend from a common ancestor with B.1.627 (bootstrap support = 99%), while B.1.634 emerges independently from the B.1 background (Fig S4); once again, the pattern is mirrored by the full S-3' phylogeny (Fig 3).

Reconciling the occurrence of the NSP6 and Orf3a deletions with the reconstruction of the evolutionary histories of these lineages for both genome segments is possible, as none of the NSP6 deletions occurred simultaneously on the same sequence, and the overlapping nature of the Orf3a deletions (i.e., FS1 is contained within FS4) make it

possible to encode ΔR1, ΔR2, FS1 and FS4 as unique traits and to map them to the phylogenetic trees (Fig 3). A third deletion, Δ69-70 on the S protein (also observed in previous VOCs and VOIs; Meng et al., 2021), was exclusive to B.1.634 and therefore not used as an informative marker. From the Orf1ab tree, ΔR1 is shared between B.1.627 and B.1.631 major, while ΔR2 is shared by B.1.631 minor, B.1.628 major and B.1.634. The most parsimonious explanation for this deletion pattern would suggest that ΔR1 occurred once (and predates the ancestral form of the Orf1ab of the B.1.627 and the B.1.631 major lineages) and that ΔR2 occurred once (predating the ancestral form of the Orf1ab for the B.1.628 major and B.1.634 lineages). Similarly, the two distinct frameshift deletions in Orf3a appear to have occurred independently: FS1 occurred once in the ancestral form of lineage B.1.634, and FS4 occurred at least once in the ancestral form of B.1.631 major, B.1.628 major and B.1.627 (Fig 3; Fig 4, upper panel). A considerable number of B.1.628 minor sequences in both trees share the “wild-type” trait (i.e., no deletion) with the Hu-1 reference genome, further suggesting that this group of sequences belongs to either B.1 or a different B.1.X lineage.

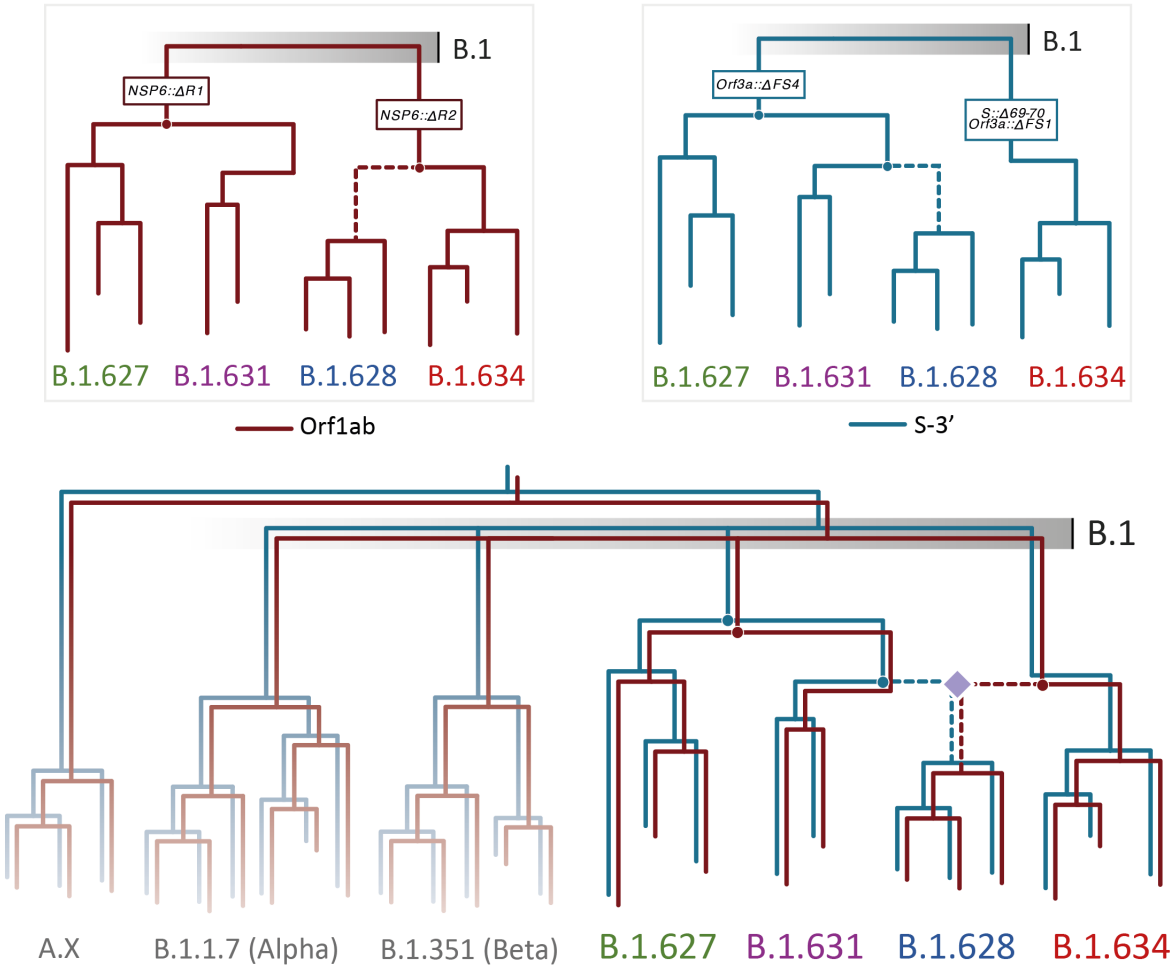


Figure 4. Schematic of the emergence of lineages B.1.627, B.1.628 major, B.1.631 major and B.1.634 from a B.1 background and their recombination history. The phylogenetic reconstruction of the evolutionary trajectories of both genome segments for the four lineages require a single occurrence of each of the deletions in ΔNSP6 and ΔOrf3a to explain most parsimoniously the observed deletion patterns (upper panel). When overlapping both phylogenies (lower panel), it can be shown that a recombination history that reconciles these deletions while maintaining the inferred ancestors through phylogenetic analyses requires the occurrence of one recombination event, where lineages B.1.631 major and B.1.634 result in the emergence of a recombinant B.1.628 major lineage. B.1.627 is evolutionarily related to B.1.631 major but not involved in the recombination event.

Mapping the deletions to a phylogenetic tree inferred for the whole genome and for the complete data set results in these appearing as homoplasic events that require

repeated occurrence; specifically, FS4 would have had to occur twice (once in the branch leading to B.1.627 and B.1.631 major and once in B.1.628 major; Fig S5). Therefore, the most parsimonious model that reconciles the minimum required number of deletions and the phylogenetic incongruencies between the evolutionary histories of both non-recombinant genome segments results in B.1.628 major having descended from a recombinant event. It inherited the Orf1ab segment (carrying the ΔR1 deletion on NSP6) from the lineage leading to B.1.634, and the S-3' segment (carrying the FS4 deletion on Orf3a) from the lineage leading to B.1.631 major (Fig 4). Visualising the SNPs of these lineages shows that B.1.628 major shares at least 6 polymorphisms with B.1.631 major in the first ~17000 nucleotides of the genome and at least 9 polymorphisms with B.1.634 in the final ~8000 nucleotides of the genome; no polymorphisms are shared between B.1.628 major and B.1.631 major along the 3' end of the genome, while no polymorphisms are shared between B.1.628 major and B.1.634 along the 5' end (Fig 5).

Figure 5. Comparison of single nucleotide polymorphisms (SNPs) between the B.1.631, B.1.628 and B.1.634 lineages. SNPs were identified in reference to the 2019 Wuhan Hu-1 reference genome (MN908947.3), shown in grey in the bottom line. Five sequences from the putative parental B.1.631 major lineage are shown on the top (purple), followed by five sequences from the putative recombinant B.1.628 major lineage on the middle (blue) and five sequences from the putative parental B.1.634 lineage at the bottom (red). Recombination point relative to the SNPs is marked by the

yellow line and by different shading colours in the genome position bar (bottom of panel).

Discussion

Genomic recombination has been widely described across sarbecoviruses in general (Boni et al., 2020) and has been identified as an important driver in the evolution of the lineage leading to the emergence of SARS-CoV-2 (Li et al., 2020; Lytras et al., 2021). More recently, recombination between the B.1.1.7 and B.1.177 lineages has been observed in the United Kingdom leading to a limited number of circulating genomes and to their designation as lineage XA, the first recombinant SARS-CoV-2 lineage under the Pango nomenclature (Jackson et al., 2021). However, no major circulating recombinant lineages spanning wider spatiotemporal distributions – across multiple countries – have been described to date. In the current work we investigate the evolutionary histories of four distinct but unusually similar SARS-CoV-2 Pango lineages circulating predominantly in the USA and Mexico and test the hypothesis that a recombination event led to the emergence of at least one of these lineages. A model can be proposed that resolves the phylogenetic incongruencies and deletion events among these lineages, in which B.1.628 major originated from a single recombination event (Fig 4). The early identification of this lineage in the USA and Mexico and its widespread circulation elsewhere represents a notable event in the COVID-19 pandemic, as no previously recognised recombinant lineages have been reported to spread across country borders and to increase in frequency at the rate observed here. Following our results, the Pango

Network committee has decided that lineage B.1.628 *major* will be designated as lineage XB, making it the second recombinant lineage in the nomenclature system.

A necessary condition for the emergence of recombinant viruses is the co-circulation of its parental lineages, as viral recombination necessarily occurs during co-infection events of a single host (Simon-Loriere and Holmes, 2011). This condition is generally observed for the four lineages under investigation, predominantly detected in the USA and Mexico (where overlapping temporal distributions suggests co-circulation), as well as in other countries in Central America (Fig 1, Table 1). The substantial number of sequences and spatiotemporal distribution of B.1.628 major might normally be interpreted as evidence of an earlier emergence compared to B.1.627, B.1.631 and B.1.634. However, sampling intensity and the relative frequency of different lineages in the region require careful consideration (Kraemer et al., 2021a), particularly given the considerable disparities in sampling intensity in the context of genomic surveillance (Brito et al., 2021).

B.1.627 and B.1.628 major were the first lineages to be detected and were particularly frequent in the USA (amongst these four lineages). Both increased in frequency in the USA between January and May 2021; whilst detection of B.1.627 declined to low levels by May, B.1.628 exhibited a second peak in detection in the USA in July 2021 (Fig 1; Fig S6). Mexico reported sequences, predominantly of B.1.628 major and B.1.634, from May to July (Fig 1; Fig S7), coinciding with the start of a systematic, nationwide genomic surveillance programme under the CoVi-Gen Mex Consortium on May 11 2021 (CONACYT, 2021). Given the differences in genome

sampling and sequencing intensity between the two countries during the months preceding the detection of these lineages (6.3% of confirmed cases were sequenced in the USA during the last week of March 2021, compared to 1.6% for Mexico; Brito et al., 2021), it is likely that early cases of the B.1.628 major were not detected by genomic surveillance. However, the regional distribution of the lineages does suggest the recombinant lineage emerged in North America between late 2020 and early 2021.

Both minor clusters identified for B.1.628 and B.1.631 fail to display the monophyly condition which defines the Pango nomenclature (Rambaut et al., 2020). The phylogenetically distinct B.1.631 minor cluster was exclusively sampled in Turkey between late June and early August. Its spatiotemporal features and phylogenetic placement (Fig 3) and its distinct genome-wide divergence to the reference genome (Fig S2) warrants further investigation. However, its relevance in the evolution of the remaining lineages appears inconsequential.

The inferred recombination breakpoint is located near the start of the S protein reading frame, generally coinciding with previously described recombination hotspots for other coronaviruses (Klerk et al., 2021; Sabir et al., 2016; Yang et al., 2021) and for SARS-CoV-2 (Boni et al., 2020; Jackson et al., 2021; Li et al., 2020; Lytras et al., 2021). However, the exact location of the breakpoint differs depending on the exact data set and method used. The two inferred breakpoints from our data sets (i.e., leading to lineage XB) are biologically relevant, located on the S protein signal peptide or on the RBD. A breakpoint on the S protein signal peptide would produce a viral genome in which each reading frame is inherited from one of the two parental lines in its entirety, while a

breakpoint on the RBD would result in a chimeric S protein. This latter possibility remains plausible given that the breakpoint would not disrupt major functional features of the protein (such as the trimeric ACE2-binding interface) and that sequence divergence remains low between these closely related lineages. Another interesting observation is that the canonical sarbecovirus transcription regulatory sequence (TRS), a 7-nucleotide sequence that regulates the viral protein expression during cell infection (Yang et al., 2021), is located 18 nucleotides upstream of the GARD breakpoint obtained with the full data set. The TRS is also associated with viral genome replication, providing a possible mechanism driving recombination at this particular breakpoint (Yang et al., 2021). From an inferential standpoint, given the uncertainty regarding the precise location of the recombination breakpoint, exploring the individual phylogenetic tree of Orf1ab independently from the phylogeny for the remainder of the genome should adequately explain the complete evolutionary history of the lineages under investigation.

Genetic recombination can have important consequences for viral adaptation and fitness, and has been observed in many other viruses. Although mechanistically-distinct to SARS-CoV-2, recombination in HIV has led to the emergence of successful circulating recombinant forms (CRFs) associated with high prevalence in some locations (Hemelaar et al., 2006; Vuilleumier and Bonhoeffer, 2015) and it has been hypothesised that enhanced replication-associated fitness may be involved (Njai et al., 2006). A hepatitis C virus (HCV) CRF was identified in St Petersburg, Russia (Kalinina et al., 2002), and has circulated for a prolonged time and across multiple countries (Raghwani

et al., 2012). The seasonal human coronaviruses (HCoVs) 229E, HKU1, NL63 and OC43 show evidence of frequent recombination among individual genome sequences and occasionally among entire clades; recombinant monophyletic clusters have been described for HCoV-OC43 and HCoV-NL63, for example (Pollett et al., 2021). This pattern extends beyond the human coronaviruses. Lineage 5 of the zoonotic MERS-CoV shows evidence of having emerged through recombination and was associated with multiple human cases in Saudi Arabia and South Korea in 2014, as well as with camel infections (Sabir et al., 2016). Given these instances, it is possible that additional undetected recombination events for SARS-CoV-2 have taken place; however, if these follow the pattern observed for B.1.628, it is likely that such lineages have also become extinct and have therefore not contributed to the dominant lineages thus far.

On the other hand, it is also possible that the widespread circulation of the B.1.628 major lineage was in part a consequence of the effects of recombination on virus fitness. However, there is no direct support for this hypothesis and the expansion of any given lineage is likely driven by a myriad of factors in addition to virus genetics (Kraemer et al., 2021b). The persistence and spread of B.1.628 major means that, at the very least, recombination had no detrimental effects on its fitness. Recombination can increase viral genetic diversity by bringing together new combinations of circulating mutations into a single genome or haplotype; this can potentially purge deleterious mutations and overcome clonal interference (Simon-Loriere and Holmes, 2011). Through this mechanism, viruses can achieve large “jumps” in sequence space without requiring the generation of intermediate forms through cumulative mutation. This is of particular

importance if these intermediate forms are selectively deleterious (Moradigaravand and Engelstädter, 2012); under such circumstances, recombination can enable virus species jump from one fitness peak to another across a valley in the fitness landscape (Crona, 2018). Exploring the extent to which recombination can drive adaptation in SARS-CoV-2 and other human coronaviruses is paramount to evaluate the long-term effects on virus evolution.

While our model for the recombinant origin of B.1.628 major reconciles the deletion and phylogenetic patterns observed in the genomic data, it still doesn't resolve all differences between the tree topologies for the Orf1ab and S-3' regions; for example, individual sequences with no deletions are occasionally found in clusters/lineage that are characterised by those deletions. If these sequences are correct, then that would imply repeated reversion of the deletion; this is thought to be highly unlikely as there are no known mechanisms by which a specific combination of nucleotides (i.e., the ancestral sequence) would be inserted into a site where a deletion previously occurred. The insertion of predictable short sequences has only been generally described for specific genetic elements (Sehn, 2015). We therefore conclude that these apparent reversions are more likely artifacts deriving from sequencing or assembly errors; this would be consistent with the presence of genome sequences in our data set where the deletion site falls among highly ambiguous positions (https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/12).

We find that B.1.628 major arose from a recombination event between the B.1.631 major clade and lineage B.1.634, prompting its designation as a recombinant lineage

under the Pango nomenclature convention. We also recommend that the group of sequences identified here as B.1.628 minor should be revisited, and potentially designated as lineage B.1. At the time of writing, the dominance of lineage B.1.617.2 (VOC Delta) appears unchallenged at a global scale, yet the remarkable expansion of B.1.628 during early- and mid-2021 highlights the viability of a recombinant SARS-CoV-2 lineage and delineates yet another key function of active genomic surveillance programmes. Our findings also emphasise the importance of further investigating the recombination rate and potential of the virus, and of exploring the drivers of such evolutionary processes.

References

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. *4*, 50–60.
Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.Y., Perry, B.W., Castoe, T.A., Rambaut, A., and Robertson, D.L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. *5*, 1408–1417.
Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Hill, S.C., Danish Covid-19 Genome Consortium, Sabino, E.C., Pybus, O.G., et al. (2021). Global disparities in SARS-CoV-2 genomic surveillance. MedRxiv Prepr. Serv. Heal. Sci. 1–14.
Callaway, E. (2021). Heavily mutated Omicron variant puts scientists on alert. Nature *600*.
CONACYT, C.N. de C. y T. (2021). Programa de Vigilancia Genómica del SARS-CoV-2 realizado por el CoViGen-Mex. *2*.
Crona, K. (2018). Recombination and peak jumping. PLoS One *13*, 1–21.
Gribble, J., Stevens, L.J., Agostini, M.L., Anderson-Daniels, J., Chappell, J.D., Lu, X., Pruijssers, A.J., Routh, A.L., and Denison, M.R. (2021). The coronavirus proofreading exoribonuclease mediates extensive viral recombination. PLoS Pathog. *17*, 1–28.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.
Hemelaar, J., Gouws, E., Ghys, P.D., and Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. Aids *20*, 13–23.
Jackson, B., Boni, M.F., Bull, M.J., Colleran, A., Colquhoun, R.M., Darby, A.C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J.T., et al. (2021). Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell 1–10.
Kalinina, O., Norder, H., Mukomolov, S., and Magnius, L.O. (2002). A Natural Intergenotypic Recombinant of Hepatitis C Virus Identified in St. Petersburg. J. Virol. *76*, 4034–4043.
Klerk, A. De, Swanepoel, P., Lourens, R., Zondo, M., Abodunran, I., Maclean, O.A., Robertson, D., Pond, S.L.K., Zehr, J.D., Stanhope, M.J., et al. (2021). Conserved recombination patterns across coronavirus subgenera.
Kraemer, M.U.G., Scarpino, S. V., Marivate, V., Gutierrez, B., Xu, B., Lee, G., Hawkins, J.B., Rivers, C., Pigott, D.M., Katz, R., et al. (2021a). Data curation during a pandemic and lessons learned from COVID-19. Nat. Comput. Sci. *1*, 9–10.
Kraemer, M.U.G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J.T., Baele, G., Parag, K. V., Battle, A.L., Gutierrez, B., et al. (2021b). Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science (80-.). *373*, 889–895.
Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature *581*, 215–220.

Li, X., Giorg, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020). Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv. *6*, 1–12.
Lytras, S., Hughes, J., Xia, W., Jiang, X., and Robertson, D.L. (2021). Exploring the natural origins of SARS-CoV-2 in the light of recombination. BioRxiv 2021.01.22.427830.
Meng, B., Kemp, S.A., Papa, G., Datir, R., Ferreira, I.A.T.M., Marelli, S., Harvey, W.T., Lytras, S., Mohamed, A., Gallo, G., et al. (2021). Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep. *35*.
Martin, D., and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563
Martin, D.P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4 : detection and analysis of recombination patterns in virus genomes. Virus Evol. 1, vev003. https://doi.org/10.1093/ve/vev003.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., and Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534.
Moradigaravand, D., and Engelstädter, J. (2012). The Effect of Bacterial Recombination on Adaptation on Fitness Landscapes with Limited Peak Accessibility. PLoS Comput. Biol. *8*, 35–37.
Njai, H.F., Gali, Y., Vanham, G., Clybergh, C., Jennes, W., Vidal, N., Butel, C., Mpoudi-Ngolle, E., Peeters, M., and Ariën, K.K. (2006). The predominance of Human Immunodeficiency Virus type I (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology *3*, 1–11.
O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al. (2021). Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. Virus Evol. *7*, 1–9.
Paradis, E., and Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.
Petrova, V.N., and Russell, C.A. (2018). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. *16*, 47–60.
Pollett, S., Conte, M.A., Sanborn, M., Jarman, R.G., Lidl, G.M., Modjarrad, K., and Maljkovic Berry, I. (2021). A comparative recombination analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. Sci. Rep. *11*, 1–11.
Raghwani, J., Thomas, X. V., Koekkoek, S.M., Schinkel, J., Molenkamp, R., van de Laar, T.J., Takebe, Y., Tanaka, Y., Mizokami, M., Rambaut, A., et al. (2012). Origin and Evolution of the Unique Hepatitis C Virus Circulating Recombinant Form 2k/1b. J. Virol. *86*, 2212–2220.
Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407.

Sabir, J.S.M., Lam, T.T.Y., Ahmed, M.M.M., Li, L., Shen, Y., Abo-Aba, S.E.M., Qureshi, M.I., Abu-Zeid, M., Zhang, Y., Khiyami, M.A., et al. (2016). Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science (80-.). *351*, 81–84.
Sehn, J.K. (2015). Insertions and Deletions (Indels). Clin. Genomics 129–150.
Sekizuka, T., Itokawa, K., Saito, M., Shimatani, M., Matsuyama, S., Hasegawa, H., Saito, T., and Kuroda, M. (2021). Genome Recombination between Delta and Alpha Variants of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). MedRxiv Prepr. Serv. Heal. Sci.
Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance *22*, 2–4.
Simon-Loriere, E., and Holmes, E.C. (2011). Why do RNA viruses recombine? Nat. Rev. Microbiol. *9*, 617–626.
Song, H., Giorgi, E.E., Ganusov, V. V., Cai, F., Athreya, G., Yoon, H., Carja, O., Hora, B., Hraber, P., Romero-Severson, E., et al. (2018). Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. Nat. Commun. *9*.
Technical Advisory Group on SARS-CoV-2 Virus Evolution (2021). Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern.
VanInsberghe, D., Neish, A.S., Lowen, A.C., and Koelle, K. (2021). Recombinant SARS-CoV-2 Genomes Circulated at Low Levels Over The First Year of The Pandemic. Virus Evol. *7*, 1–12.
Vuilleumier, S., and Bonhoeffer, S. (2015). Contribution of recombination to the evolutionary history of HIV. Curr. Opin. HIV AIDS *10*, 84–89.
Xiao, Y., Rouzine, I.M., Bianco, S., Acevedo, A., Goldstein, E.F., Farkov, M., Brodsky, L., and Andino, R. (2016). RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence. Cell Host Microbe *19*, 493–503.
Yang, Y., Yan, W., Hall, A.B., and Jiang, X. (2021). Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination. Mol. Biol. Evol. *38*, 1241–1248.

5 Spatio-temporal trends of arboviral spread in Mexico and Central America

Highlight: Comparing the spatial dynamics of viruses transmitted by the same vector can help overcome *limited sampling*

Dengue viruses (DENV) circulate seasonally in Mexico, while Chikungunya virus (CHIKV) caused a single large epidemic following its introduction into the Americas during the early and mid-2010s. In this chapter we explore the spatial dynamics of DENV-1, DENV-2 and CHIKV over different timescales and find similar patterns for all three viruses, as they are introduced repeatedly into the same region in the south of the country. The improved genomic representation of CHIKV compared to DENV-1 and DENV-2, despite the limited sampling time range, allow us to formulate hypotheses on the drivers of DENV circulation within the region.

Abstract

Established and emerging arboviral epidemics are common in the Americas and comprise both seasonal epidemics (e.g. dengue virus, DENV), and large emerging outbreaks (e.g. Chikungunya and Zika viruses, CHIKV and ZIKV). In Mexico and Central America these viruses share the same geographic and demographic context, and also share a mosquito vector species, *Aedes aegypti*, hence a comparative analysis of how they spread through space and time may provide new insights into their

transmission. Here we generate new CHIKV, DENV-1 and DENV-2 genetic sequences and analyse them in combination with publicly available genomic data from Central America in order to understand the geographical and temporal patterns of arboviral circulation in the region. Our new data expands the available genomic data for CHIKV in Mexico from 8 to 39 complete genome sequences. Phylogenetic analyses show that the three viruses descend from lineages that circulated predominantly in the Caribbean and Central America. Discrete phylogeographic analyses show that multiple introductions to Mexico are a shared feature of these arboviruses; CHIKV was introduced at least on 12 occasions over a year, with six introductions leading to extensive ongoing transmission within Mexico. Further, both DENV serotypes were introduced at least seven times over a decade. For both CHIKV and DENV-2, the southwest Mexico is the most frequent inferred ancestral location for virus transmission lineages within Mexico. Both viruses spread within the country, most prominently across the Pacific coast into the central western region, and into the south-eastern Yucatán peninsula. When integrating epidemiological data from CHIKV and DENV in Mexico, it becomes apparent that new introductions lead to increases in numbers of cases. These phylogeographic results are also interpreted in the context of cross-border anonymised human mobility data. The Southeast region of Mexico is likely less affected by arbovirus transmission dynamics in other Mexican regions, possibly given its high connectivity with Belize. Our results suggest these three arboviruses in Mexico share similar transmission patterns and reveal the pivotal role of the Southwest in the seeding and spread of arboviruses across the country.

Introduction

The spread of emerging arboviruses is driven by complex interactions among environmental conditions *(1, 2)*, ecological factors that affect reservoir and vector species *(3, 4)*, human behaviour and mobility *(5–10)* and previous protective population immunity in the host population *(11–13)*. Untangling the individual contributions of these determinants across spatio-temporal scales is challenging. While some arboviruses display seasonal dynamics, with varying transmission intensity across seasons *(14–17)*, novel or re-emerging arboviruses can cause large outbreaks in locations where populations have limited or non-existent prior immunity to the virus *(18–22)*. However, comparative study of the common features of arbovirus transmission dynamics can provide powerful insights into their establishment and transmission dynamics in different populations. For example, arboviruses transmitted by the same vector species (e.g., *Aedes aegypti*) are expected to exhibit shared trends in spatio-temporal patterns. These similarities could provide valuable information to better understand the transmission dynamics of current and future arboviruses. Further, mapping the emergence and spread of multiple arboviruses can inform the development of efficient spatially targeted interventions.

An interesting comparison can be established from the emergence of Zika virus (ZIKV) and Chikungunya virus (CHIKV) in the Americas, a region where seasonal arboviruses (most notably Dengue virus, DENV) circulate endemically and cause large outbreaks. While these viruses species belong to different families (ZIKV and DENV to the Flaviviridae, and CHIKV to the Togaviridae), they are all transmitted primarily by

Aedes aegypti mosquitoes. DENV and CHIKV can be also transmitted by *Aedes albopictus* mosquitoes but this species is believed to be a secondary vector *(23)*, an observation that has been speculated for ZIKV as well *(24, 25)*. It should be therefore possible to gain insights into the transmission dynamics of viruses that share a transmitting vector and, to an extent, an ecological niche. The extent to which observations among viruses can be extrapolated will be however limited by the presence of other factors, such as previous immunity and cross-immunity *(26–28)*.

Barring the possibility that some 17th and 19th century cases of dengue fever in the Americas were in reality caused by CHIKV *(29, 30)*, autochthonous chikungunya fever cases had not been reported in the continent until the mid 2000s *(31)*. CHIKV emerged in the region and caused large outbreaks from 2013 to 2017. The first locally-acquired CHIKV infections were reported in the Caribbean islands in 2013 *(32, 33)* and spread in 2014 to Central and South America *(34–36)*. Most outbreaks have been attributed to a single introduction of the Asian genotype into the Eastern Caribbean *(37)*, except for Brazil, where a second independent introduction of the Eaastern/Central/South African (ECSA) genotype was identified *(38)*. The ECSA genotype has since spread across Brazil *(39, 40)* and was recently introduced to Paraguay *(41)*.

The CHIKV Asian lineage outbreak that circulated among various Caribbean territories over 2014 *(42)* caused high incidence in Guadeloupe (295.8 cases per 1000 inhabitants in November 2014) and French Guiana (112.41 cases per 1000 inhabitants in July 2014), and large numbers of cases in Dominican Republic (over 142 000 new

cases reported in July 2014). CHIKV reached South America and the USA in June 2014, and reached various countries in central America and Mexico by the end of that year *(43)*. The spread of CHIKV across the Americas was at least partially driven by international travel *(44–46)*. The emergence of CHIKV was highly likely given (i) the absence of prior immunity to CHIKV in the Americas, (ii) the presence of both *Ae. aegypti* and *Ae. Albopictus* in the region, and (iii) high travel volumes from countries with significant levels of autochthonous virus transmission *(31)*.

CHIKV surveillance in Mexico commenced formally in 2014 under a National Epidemiological Surveillance System ('Sistema Nacional de Vigilancia Epidemiológica – SINAVE'). During 2015 a high number of CHIKV cases were reported across the country, particularly in some southern and southwestern states such as Guerrero, Michoacán and Yucatán *(47)*. Following an explosive outbreak of CHIKV during 2015 that peaked in July-August, reported case numbers rapidly decreased towards the end of 2015. From 2016 onwards, limited numbers of cases have been reported and limited to specific regions, although it has been suggested that CHIKV incidence in southwestern states like Guerrero could be underestimated *(48)*. The available virus genomic data from Mexico suggests that multiple introductions have occurred *(49–52)*, but the domestic dissemination of CHIKV from these introductions has not been explored.

The dynamics of DENV in the region are somewhat different. While reports of Dengue-like illness and dengue fever outbreaks have been reported in the Americas since the late 1700s *(53)*, a considerable epidemiological shift during the 1950s and

1960s saw the establishment of large and more frequent outbreaks across the region *(54)*. By the 1980s and 1990s, DENV outbreaks were being reported yearly in at least 24 countries in the Americas, a pattern potentially driven by the re-establishment of previously eradicated *Ae. aegypti* populations *(54)*. DENV is divided into four immunologically and genetically distinct serotypes named DENV-1, DENV-2, DENV-3 and DENV-4. Before the 1960s, DENV-2 was the only serotype detected in the Americas but reports of the emergence of the other three serotypes were published during 1960s to 1990s. This included multiple introductions of genetically distinct DENV-2 lineages *(53, 55)*.

Once established, DENV epidemics follow seasonal patterns that are now common in several American countries *(10, 56–59)* and particularly evident in regions where the virus is endemic (i.e., where at least one serotype is constantly present; *60*) or hyperendemic (i.e., where multiple DENV serotypes co-circulate). In these settings, yearly seasonal cycles are characterised by periods where outbreaks are more frequent, driven in part by climatic factors *(61, 62)* which improve the conditions for vector breeding and increase vector-host interactions *(63, 64)*.

Beyond these seasonal patterns, the co-circulation of different DENV serotypes can lead to specific viral types becoming dominant across multiple years *(65, 66)*. The processes driving long-term DENV serotype dynamics are complex and include interactions among factors relating to climate, demography and population immunity *(11, 67, 68)*. In many countries in the Caribbean, South and Central America, DENV

follows a seasonal pattern in which large outbreaks occur during the rainy seasons *(10, 56–59)* that vary in size (in terms of the number of cases of classic dengue fever, DF, and of dengue haemorrhagic fever, DHF).

As in other regions in the Americas, DF and DHF surveillance in Mexico has been ongoing since the 1980s when DENV-1, DENV-2 and DENV-4 co-circulated persistently *(69)*; DENV-3 was first detected in the country in 1995 *(70)*. Dengue rose to prominence at the national level when the incidence of DHF increased during the 1990s, associated with the re-discovery of *Aedes* mosquitoes in Mexico which were possibly re-introduced into the region *(69)*. The most affected areas have consistently been the southern and coastal regions: the southern border states of Chiapas and Oaxaca, the southwestern state of Guerrero, the states located in the Yucatán peninsula (Yucatán, Quintana Roo and Campeche), and the Gulf Coast states of Veracruz and Tamaulipas *(69, 70)*. This geographical distribution marginally expanded during the 2000s into locations closer to central Mexico, such as the state of Morelos and the higher altitude areas of Guerrero *(71, 72)*.

Hotspots of Dengue fever (and other arboviral diseases) in Mexico during the early 2010s were identified in and around urban areas in the southern states, where transmission occurs year-round but peaks during the rainy season *(72)*. DENV serotype incidence fluctuates over multiple years; during the early 2000s, DENV-1 and DENV-2 alternated as the dominant serotype in Mexico, with DENV-1 replacing the previously dominant DENV-2 between 2004 and 2007 *(71)*. Interestingly, the introduction of

CHIKV and ZIKV in the country coincided with a decrease in DENV incidence in endemic states during the peak of both epidemics *(73)*. Genomic analyses of DENV in Mexico (sampled from both humans and mosquitoes; *(74)*) suggest that lineage replacement is a frequent feature of viral circulation in the country *(75, 76)*, a pattern similar to that observed in southern China and Southeast Asia *(77)*. However, the dynamics and effects of these lineage replacements have not been formally evaluated in Mexico.

In this study, we present a new set of 39 CHIKV, 7 DENV-1 and 11 DENV-2 genome sequences from Mexico, obtained from a set of historical samples (collected between 2013 and 2017 from across the country), that form part of the SINAVE national sentinel arbovirus surveillance program. We use spatial phylodynamic methods to quantify and compare the patterns of lineage importations and spread of all three viruses into different regions of Mexico. By interpreting our findings alongside epidemiological and human mobility data, we find similarities in the spatial epidemiology of these viruses and explore the extrinsic factors that may drive arbovirus establishment and seasonality in Mexico. We find striking similarities between the introduction of DENV and CHIKV into Mexico across different timescales and conclude that human mobility may not entirely account for virus movements in the region.

Methods

DENV and CHIKV genome sequences

The *National Arbovirus Reference Laboratory* at the National Institute for Epidemiological Diagnosis and Reference (Instituto de Diagnóstico y Referencia Epidemiológicos "Dr. Manuel Martínez Báez" – InDRE, Ministry of Health of Mexico and PAHO partner) in Mexico City periodically processes and stores samples derived from confirmed CHIKV and DENV diagnosed cases, as part of the National Sentinel Program for Arbovirus Epidemiological Surveillance *(78)*. Blood samples suspected of arboviral febrile illness that test positive in state laboratories across the country are posted to InDRE to be processed and stored at -80°C for long-term storage, resulting in a historic collection of samples spanning multiple years.

We generated a small batch of whole genome sequences for Dengue virus 1 (DENV-1), Dengue virus 2 (DENV-2) and Chikungunya virus (CHIKV) collected from human and mosquito samples using a primer scheme approach *(79)* and Oxford Nanopore sequencing protocols. Candidate samples for sequencing were pre-selected from those with with sampling dates and locations for which publicly available genomic data from Mexico was scarce or non-existent; samples were also selected by viral RNA concentration (diagnostic qPCR $C_t \leq 35$). From these pre-selected samples, we randomly chose samples to sequence so that years and states with low or no representation in public data bases were prioritised. Total RNA was extracted and used to generate cDNA using the SuperScript™ IV First-Strand Synthesis System (Invitrogen, CA, USA) and random hexamer primers. Following this, cDNA were subject to multiplex PCR

reactions using virus-specific primer sets *(39, 80)* that generate ~400 nt partially overlapping PCR fragments covering the entirety of the viral genome. PCRs were performed using the Q5 High-Fidelity DNA polymerase kit (New England Biolabs, MA, USA) and run for 35-40 cycles. PCR products were purified using AmpureXP magnetic beads (Beckman Coulter, CA, USA) and quantified on a Qubit 3.0 spectrophotometer device using the Qubit dsDNA High Sensitivity kit (Life Technologies, CA, USA). Sequencing libraries were prepared from the purified PCR products by adding barcoding fragments to individual samples using the EXP-NBD104, EXP-NBD114, and SQK-LSK109 kits (Oxford Nanopore Technologies, UK). Sample concentrations were then normalised and pooled into equimolar mixes and loaded onto R9.4 flow cells (Oxford Nanopore Technologies, UK). Sequencing runs were carried out until sufficient genome coverage was achieved as monitored in real-time using RAMPART (https://artic.network/ncov-2019/ncov2019-using-rampart.html). Individual genomes were reference assembled, and their assignment to specific serotypes or genotypes was performed using the Arbovirus Typing Tools from the Genome Detective web server *(81)*.

Phylogenetic analyses of DENV-1, DENV-2 and CHIKV in the Americas

We compiled a collection of whole genome sequences (i.e., sequences > 10000 nt in length for DENV and > 12000 nt in length for CHIKV) for DENV-1 (Genotype V), DENV-2 (Genotype III) and CHIKV (Asian Genotype) by retrieving available sequences in GenBank (as of 2020-02-01, the date predating the selection and

sequencing of the new samples) that were collected from any country in North, Central and South America or the Caribbean. We further expanded this set of sequences by adding 16 DENV-1 and 14 DENV-2 complete genomes collected in Nicaragua between 2013 and 2019 from a paediatric cohort *(59, 82)*. Each data set was aligned using *MAFFT (83)* and visually inspected for accuracy. Alignments were trimmed to remove the untranslated terminal regions (UTRs), producing final alignments of 10180 nucleotides for the DENV-1, 10023 nucleotides for DENV-2 and 11151 nucleotides for CHIKV.

Maximum likelihood (ML) phylogenetic trees were estimated using *IQtree 2.0 (84)* for each data set. Trees were inferred under a General Time Reversible (GTR) substitution model, and substitution rate heterogeneity across sites in the alignment was modelled using a Gamma distribution with four discrete categories. Node support on all phylogenies was estimated using non-parametric Shimoda-Hasegawa approximate Likelihood Ratio Tests (SH-aLRTs) through 1000 replicates *(85)*. The resulting trees were midpoint rooted and annotated with the sequence collection country and region.

To evaluate the temporal signal within our data sets, we plotted regressions between sample root-to-tip genetic distance and sample collection date, using *TempEst v1.5.3 (86)*. Correlation coefficients were used to assess the degree of "clock-like" evolution exhibited by each data set.

Phylogeography of the Central American DENV-1, DENV-2 and CHIKV lineages

The initial phylogenies for each virus showed that sequences from Mexico clustered into single, well-supported phylogenetic lineages together with other sequences from Central America, the Caribbean and sporadic sequences from elsewhere in the Americas (see *Results*). We therefore explored the spatiotemporal spread of the three viruses within these lineages using Bayesian phylogenetic methods.

We estimated time-calibrated phylogenetic trees using *BEAST v1.10.4 (87)* independently for each virus, using the collection date for each sequence. In two cases (for DENV-1) where no collection date was available, a uniform prior was assigned to the corresponding tip. These analyses were performed using a Hishino-Kasegawa-Yang (HKY) substitution model with rate heterogeneity modelled as a Gamma distribution. A Skyride coalescent tree prior *(88)* was used for all data sets, and a relaxed uncorrelated clock model *(89)* was used to account for rate variation across lineages within the phylogeny, with a continuous-time Markov chain (CTMC) rate prior *(90)*. *BEAST* explores the tree space given a model and a set of priors using a Markov Chain Monte Carlo (MCMC) algorithm; the MCMC chains were run by duplicate for 100 000 000 MCMC steps, with the first 10 000 000 steps discarded as a burn-in period. The results from the independent runs were combined using *LogCombiner,* resulting in an empirical distribution of 35 000 trees which were summarised as a Maximum Clade Credibility (MCC) tree using *TreeAnnotator*. Convergence for individual MCMC chains (defined as all relevant parameters achieving effective sample size, ESS, values >200) was

assessed with *Tracer (91)*; parameters that did not reach convergence are further discussed in the *Results* section.

The spatial spread of individual lineages between countries was explored through a discrete phylogeographic analysis (discrete trait analysis, DTA; *(92)*). The tips were assigned either to the country where the sequences were collected or to the geographic region containing the country of origin. The previously generated sample of posterior trees was then used as an empirical tree distribution that was resampled during a Bayesian reconstruction of ancestral node locations, using an asymmetric location trait model *(92)*. Transition rates between locations that significantly explain the diffusion process for each phylogenetic tree were identified by performing Bayes Factor (BF) tests of the posterior odds identifying said rates as non-zero, and explored using the Bayesian Stochastic Search Variable Selection (BSSVS) approach implemented in *BEAST (92)*. We consider that BF > 4 defines well-supported diffusion rates and plot the higher 85th percentile of all transition rates (posterior probability, PP > 0.5). The expected number of DTA transitions between countries and regions were estimated using a robust counting approach *(93)*. As before, an MCC tree was generated to summarise the posterior tree sample and estimate the posterior probabilities for the locations of the tree nodes.

Historical epidemiological and human mobility data

Individual case information from patients with laboratory-confirmed dengue fever (DF) or Chikungunya infection is routinely collected by the National Institute for

Epidemiological Diagnosis and Reference (InDRE) as part of the national epidemiological surveillance programme. We collected individual information of confirmed cases between 2010 and 2019 including, clinical (date of symptom onset, date of sample collection, final diagnostic), demographic (patient age and sex) and geographic data (collection state, municipality and locality) associated with each individual case. A portion of the cases determined as positive for DF also include the associated DENV serotype, identified through serology or PCR product sequencing.

To further incorporate the effects of connectivity between states and human mobility into our analyses, we obtained human mobility data derived from de-identified smartphone usage from devices whose users opted in to the Google Location History feature (https://support.google.com/accounts/answer/3118687) as described by *(94)*. The data was collected between 2019-10-27 and 2020-02-29 to ensure that mobility estimates were not affected by the implementation of non-pharmaceutical interventions (NPIs) during the COVID-19 pandemic *(95)*. Relative flow (f_r) estimates between location pairs (a,b) within calendar weeks (t) were gathered and aggregated for individual Mexican municipalities and at the country-level for the neighbouring Central American countries of Belize, El Salvador, Guatemala, Honduras and Nicaragua. These f_r estimates are already normalised by the maximum flow between location pairs for a given week t', such that $f_r(a,b,t) = f(a,b,t)/max_{i,j,t}f(i,j,t')$ where (i,j) correspond to the locations showing the highest flow at week t'. The Mexico data was further grouped by state, and all f_r estimates between municipalities between states at every week t were averaged. Weekly mobility estimates are variable, particularly by late December 2019.

We therefore grouped our observations into bi-weekly periods to visualise the variation of mobility trends across time. Median values for f_r estimates during these time periods were calculated and normalised relative to the maximum value for our entire data set.

Results

CHIKV, DENV-1 and DENV-2 genome sequences from historical samples from Mexico

CHIKV samples were selected to represent isolates from 18 out of the 32 Mexican federal states, collected between May 2014 to December 2015 which covers the entire CHIKV epidemic in the country. From these, we generated 39 complete CHIKV genomes sequences that were subsequently assigned to the Asian genotype; this is the main genotype circulating in the Americas *(96)*. These new sequences add substantially to the 8 previously available complete CHIKV viral genomes from Mexico, generated from targeted investigations or opportunity sampling. Our data also increases the range of available CHIKV genomes to 16 additional federal states for which no CHIKV genomes were available (previous sequences were mainly from Chiapas - 6 sequences, Tamaulipas - 1 sequence and Jalisco - 1 sequence). The sampling timespan for CHIKV in Mexico has been also expanded: previously available sequences were collected during the start of the Mexico epidemic between October and November 2014 (7 sequences) and during November 2015 (1 sequence). The new sequences therefore represent the entirety of the 2015 CHIKV epidemic in Mexico that peaked around July-August 2015 (Fig 1B; Table 1).

Table 1. Arbovirus cases and genomes across geographic regions in Mexico, including the sequences generated in this study.

Region	State	CHIKV		DENV-1		DENV-2	
		Cases[a]*	Seqs[c]	Expected cases[b]**	Seqs[c]	Expected cases[b]**	Seqs[c]
Northwest	Baja California	905	2	16703	--	9522	--
	Baja California Sur						
	Chihuahua						
	Durango						
	Sinaloa						
	Sonora						
Northeast	Coahuila	240	4	6618	--	14207	--
	Nuevo León						
	Tamaulipas						
Center-north	Aguascalientes	31	3	6749	--	314	--
	Guanajuato						
	Querétaro						
	San Luis Potosí						
	Zacatecas						
Center-south	Mexico state	795	6	5613	1	1150	1
	Mexico City						
	Morelos						
West	Colima	2959	7	20145	3	8839	4
	Jalisco						
	Michoacán						
	Nayarit						
East	Hidalgo	2501	6	17267	1	10225	4
	Puebla						
	Tlaxcala						
	Veracruz						
Southwest	Chiapas	3951	12	9994	1	16150	6
	Guerrero						
	Oaxaca						
Southeast	Campeche	2276	7	12839	70	6384	11
	Quintana Roo						
	Tabasco						
	Yucatán						

*Source: SINAVE data, InDRE/Ministry of Health, Mexico.
**Calculated as the relative proportion of the serotype relative to all serotyped cases multiplied by the total of reported DF/DHF cases, rounded up. Source: SINAVE data, InDRE/Ministry of Health, Mexico.
[a]Reported between 2015 and 2018.

[b]Reported between 2013 to 2018.
[c]All publicly available complete genome sequences used in this study.

We analysed DENV-positive samples and mosquito-pooled samples from six southern states, collected between 2013 and 2017. We successfully generated 7 complete DENV-1 and 11 DENV-2 partial genomes. For DENV-2, there was insufficient genome coverage for a number of genome regions, including both UTRs, and therefore these regions were masked as Ns. Nonetheless, 5104/10177 sites (~50% of the genome length) showed adequate coverage (>20x) providing sufficient signal for phylogenetic analyses within the context of other publicly available viral sequences from other regions of Mexico and the Americas *(97)*. The DENV-1 sequences we obtained were from the states of Quintana Roo (1), Jalisco (3), Veracruz (1), Chiapas (1) and Morelos (1), with sampling dates from 2013 to 2017; the DENV-2 sequences were from the states of Chiapas (3), Jalisco (3), Quintana Roo (1), Colima (1), Veracruz (2) and Morelos (1), with sampling dates between 2013 and 2016. Our sequencing data represent a significant improvement in the spatio-temporal representation of DENV genome sequences from Mexico, which previously accounted for sequences up to 2012 for DENV-1 and 2011 for DENV-2. Furthermore, two previously unrepresented states feature in our DENV-1 sequences (Chiapas and Veracruz) and four previously unrepresented states for DENV-2 (Colima, Jalisco, Morelos and Veracruz; Table 1).

CHIKV epidemiological trends in the Americas and Mexico

The CHIKV epidemic in Mexico started approximately one year after the first locally acquired infections were reported in the Caribbean islands in late 2013. As

previously described *(32, 33)*, the first cases were reported in the Eastern Caribbean (also named the Lesser Antilles; Table S1) and likely sparked the first large epidemics in other Central Caribbean islands; Fig 1A). During these initial waves, the Dominican Republic (Central Caribbean) and Guadeloupe (Eastern Caribbean) reported over 100,000 new monthly cases at their respective peaks, around July (over 142,000 cases in the Dominican Republic) and November (over 118,000 cases in Guadeloupe; Fig S1A). Central America and the Andean region of South America saw large epidemics during 2015 and early 2016 but were generally less severe than those in the Caribbean (Fig 1A). During this time, Colombia reported the largest number of cases for a single country in the region, reporting over 55,000 cases during their epidemic peak in March 2015 (Fig S1A). The southern cone region of South America drove a resurgence of CHIKV cases in the continent over the following year, between March 2016 and January 2017 (Fig 1A), dominated by infections in Brazil (Fig S1A; *(34–40)*).

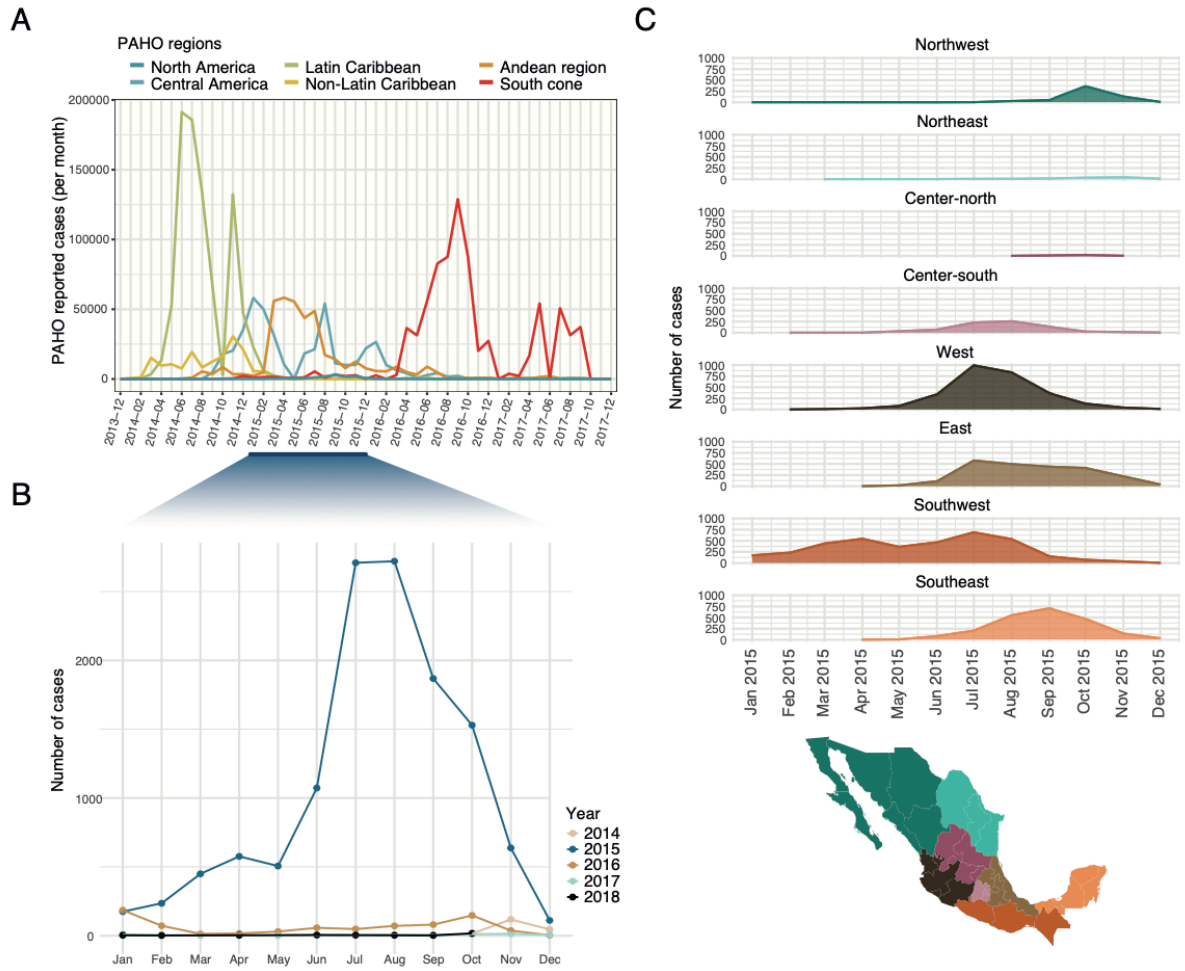


Figure 1. CHIKV epidemiological trends in the Americas and Mexico from 2013 to 2018. ***(A)*** Monthly number of CHIKV cases reported to the Pan-American Health Organisation (PAHO) between 2013 and 2017, grouped by PAHO region. The largest epidemic peak in Mexico occurred in 2015, highlighted in blue. ***(B)*** Monthly number of confirmed CHIKV cases in Mexico grouped by year, reported by the SINAVE national surveillance system of Mexico. ***(C)*** Monthly number of confirmed CHIKV cases per geographic region in Mexico during 2015. A map of the states included in each geographic region is shown below. Details of the states included in each region are available in Table 1.

While Mexico reported CHIKV cases from the end of 2014 through to 2018, cases grew rapidly in January 2015 and peaked in summer 2015 (Fig 1B). New case numbers grew from 506 in May to 2706 in July and 2720 in August 2015. Spatially, cases started increasing in the Southwest region, followed by epidemic peaks in the East (Gulf of

Mexico coast) and West (Pacific Ocean coast) regions, as well as the Southeast region, which encompasses the Yucatán peninsula (Fig 1C). A small number of cases were also observed in the Center-south region, which comprises Mexico City and its surrounding areas and. Later in the year, a small peak in cases was observed in the Northwest (120 cases, relative to 1000+ weekly cases during the peak of the epidemic), which is also the only region that saw a modest second wave of CHIKV cases around January 2017 (Fig S2).

Genomic sequencing of CHIKV in the region gradually increased during 2014 but was maintained at <10 virus sequences per month (Fig S1B). Steady numbers of CHIKV genomes from the USA, various Caribbean territories and Central American countries were generated and publicly shared throughout 2014. From December 2014 through 2015, limited numbers of genomes were generated by Nicaragua and Mexico (including the 39 sequences newly generated in this study). Overall, the numbers of sequences per country over this period does not correlate with the cumulative reported cases per country (Spearman's Rho = 0.26, $p = 0.15$; Fig S3).

Phylogeography of CHIKV in the Americas and Mexico

The CHIKV epidemic in the Americas has been documented since its detection in 2013 *(34, 38, 45, 52, 98–101)*. Our ML tree shows that the lineage circulating in the Americas (SH-aLRT = 97.9) descends from sequences from Southeast Asia and the Pacific Islands, suggesting a single introduction of the CHIKV Asian genotype to the Americas *(96)*.

A rapid early expansion of CHIKV across North and South America is apparent at the base of the American clade, as sequences collected from different countries group into poorly supported clusters (Fig S4). Within the American CHIKV phylogeny there is a large lineage circulating predominantly in the Caribbean, Central America and North America (henceforth referred to as the CCNA lineage, Fig 2A), which includes all the sequences from Mexico (SH-aLRT = 77.5; Fig S4).

A time-calibrated phylogenetic analysis of the CCNA lineage indicates the date of the root of the CCNA lineage is around mid-September 2013 (Median = 2013.69, 95% HPD = 2013.59 – 2013.79; Fig 2B), consistent with previous results *(96)*. The basal branches of the CCNA lineage show early viral circulation predominantly in the Eastern Caribbean and multiple lineage exportations to various locations in the Central Caribbean, South, Central and North America, including two importations into Mexico. These account for most of the earliest sequences generated in the country, specifically, InDRE04 (a single sequence from the western state of Jalisco, *(49)*) and a cluster of sequences from the southwestern state of Chiapas *(98)*. While this basal clade (denoted here as Clade A, Fig 2B) was not sampled after late 2014 (with the exception of one sequence from Colombia), a second well-defined clade (node posterior probability, PP = 0.91) circulated in the Eastern Caribbean (Clade B, Fig 2B; Fig S4). The sequences from Mexico in this clade are descendants from viruses circulating in Haiti, Nicaragua and the USA; nonetheless, given the overrepresentation of sequences from the USA and Nicaragua (Fig S1B), it is likely that sampling bias partially accounts for the inferred

location of these internal nodes. This could be the case for Nicaragua, commonly inferred as the ancestral location from which imports into Mexico occur but which might exclude other relevant countries in the region with limited genomic coverage such as Guatemala, Honduras, El Salvador and Belize.

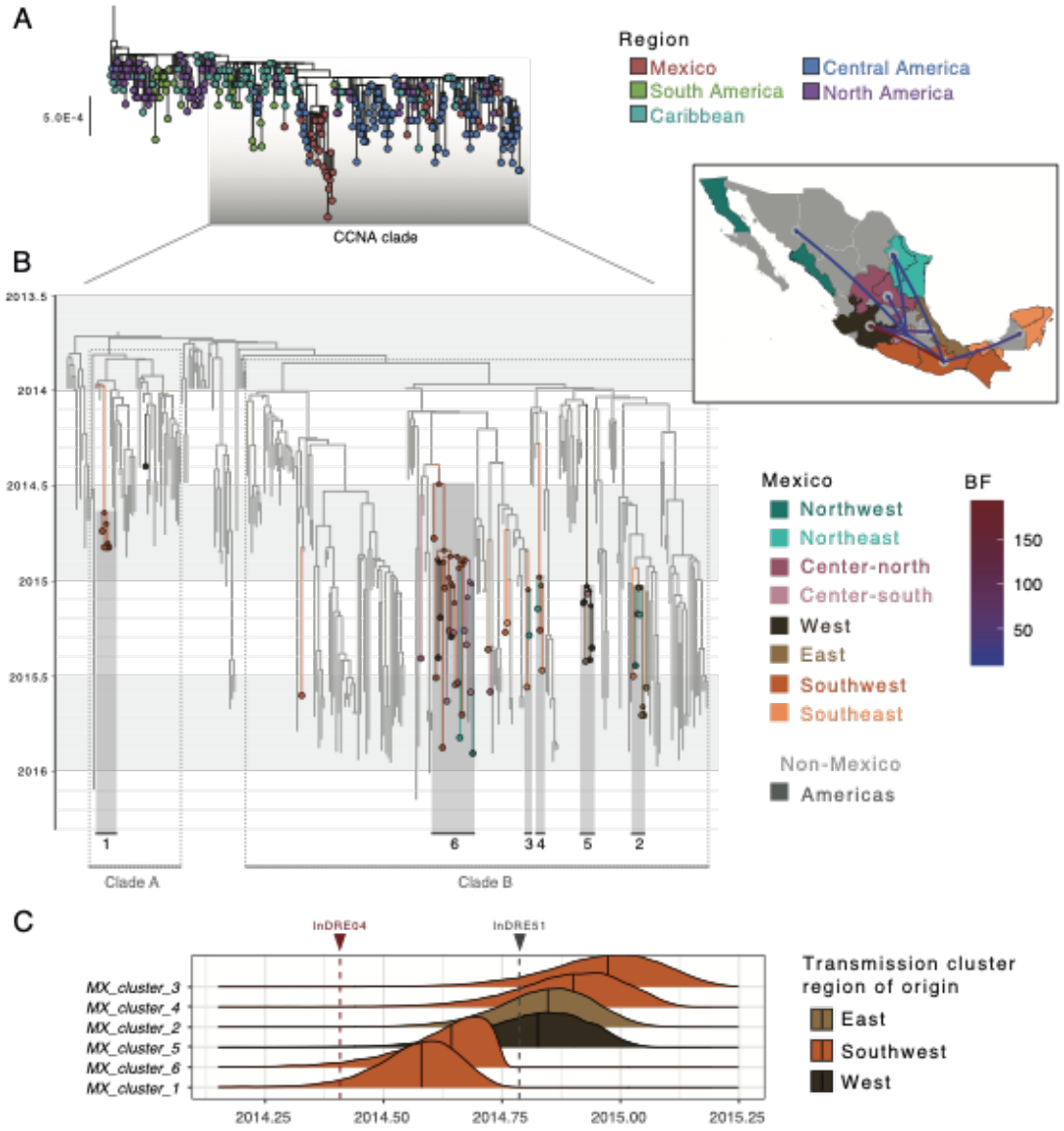


Figure 2. Importation and domestic spread of CHIKV in Mexico. ***(A)*** ML phylogenetic tree of all CHIKV complete genome sequences from the Americas included in our analysis. Tree tips are coloured by region where they were collected. Tree tips from Asia and the Pacific islands have been removed for this visualisation. ***(B)*** DTA phylogeographic analysis of the CHIKV CCNA lineage, coloured by sample

collection region in Mexico. Key, well defined clades A and B are shown, and clusters of sequences within Mexico are identified and numbered 1 to 6. Tip symbols and nodes are shown for locations within Mexico and are coloured by region. The map in the inset shows pairs of locations among which there exist significant pairwise DTA transition rates, as inferred by a BSSVS analysis. Transition rates with a posterior probability > 0.5 are shown and coloured by Bayes Factor (BF). ***(C)*** Posterior probability densities of the TRMCA of six CHIKV transmission clusters in Mexico (corresponding to the numbering in panel B). Median values of each distribution are shown for each cluster, and the distributions are coloured by the most probable location of each MRCA. The dotted lines show the collection date for the first travel-history associated CHIKV case identified in Mexico (red, InDRE04, collected on 2014-05-30 and found in Clade A) and the first confirmed autochthonous CHIKV case in the country (grey, InDRE51, collected on 2014-10-15 and found in *MX_cluster_6*).

A discrete phylogeographic reconstruction of the importations and circulation of CHIKV in Mexico shows at least 13 introductions of CHIKV lineages into the country. From a visual inspection of an MCC phylogenetic tree, seven of these did not cluster with additional sequences from within the country and represent either returning travellers (as is the case with the aforementioned InDRE04 sequence) or poorly sampled viral lineages within the country. We found least six well-supported clusters that contained only sequences from Mexico in the MCC tree (named *MX_cluster_1* to *MX_cluster_6*, Fig 2B). The MRCAs of four of the six clusters were inferred to have existed in the Southwest ($PP_{MX\_cluster\_1}$ = 0.56, $P_{Southwest}$ = 0.76; $PP_{MX\_cluster\_3}$ = 0.99, $P_{Southwest}$ = 0.66; $PP_{MX\_cluster\_4}$ = 0.95, $P_{Southwest}$ = 0.84; $PP_{MX\_cluster\_6}$ = 0.99, $P_{Southwest}$ = 0.99), East ($PP_{MX\_cluster\_2}$ = 0.54, P_{East} = 0.68) and West regions ($PP_{MX\_cluster\_5}$ = 0.50, P_{West} = 0.91; Fig 3B). The BSSVS approach, which explores the inclusion or exclusion of specific location transition rates and their likelihood in explaining the spatial diffusion process within the phylogeny, identified three significant transition rates (BF > 4) in the phylogeny between the Americas and Mexico, namely movements into the Southwest

(BF = 5236.52; PP > 0.99), Center-north (BF = 86.21; PP = 0.92) and West (BF = 15.76; PP = 0.68) regions.

The MRCAs of the six largest Mexican clusters were dated between mid- and late-2014 (Fig 2C). Two of these (*MX_cluster_1* and *MX_cluster_6*) predate the first reported autochthonous CHIKV case in the country, InDRE51 *(49)* but are more recent than the earliest reported CHIKV case in Mexico, InDRE04, for which the patient had a travel history to the Caribbean *(49)*. The remaining four clusters have circulated in the country since October 2014 (median age $MRCA_{MX\_cluster\_2}$ = 2014.85, 95% HPD = 2014.62 – 2014.04; median age $MRCA_{MX\_cluster\_3}$ = 2014.97, 95% HPD = 2014.71 – 2015.18; median age $MRCA_{MX\_cluster\_4}$ = 2014.89, 95% HPD = 2014.60 – 2015.11; median age $MRCA_{MX\_cluster\_5}$ = 2014.83, 95% HPD = 2014.59 – 2015.01).

Reconstructed viral movements inside Mexico predominantly occurred across the southern and central coastal regions, with a south-to-north trend. BSSVS identified 14 significant state transitions inside Mexico in the phylogeny, with the most predominant source regions being the West, Southwest and Center-south (Table S2; Fig 2B inset). *MX_cluster_1* comprises the largest continuous CHIKV transmission cluster in Mexico (18 sequences from Mexico and one from Nicaragua), sampled across 11 different states (Baja California, Mexico state, Mexico City, Guerrero, Tamaulipas, San Luis Potosí, Quintana Roo, Jalisco, Morelos, Michoacán and Oaxaca) and in seven out of eight regions in the country (Fig 2B). *MX_cluster_6* corresponds to a previously described early cluster of cases from the state of Chiapas in the Southwest region *(98)*. The

remaining clusters show more limited levels of circulation across regions: *MX_cluster_3* and *MX_cluster_4* were sampled only in two regions each, *MX_cluster_5* circulated in the East and West regions and *MX_cluster_2* was observed in three regions (East, Northeast and Northwest).

Epidemiology of Dengue virus 1 and 2 in Mexico, 2013-2019

Data from the SINAVE program shows clear seasonal patterns of dengue fever in Mexico between 2013 and 2019, with distinct trends among regions (Fig S5). The Northeast, East and Southeast regions of the country reported between 2000 - 3000 monthly cases at their respective 2013 peaks, while a less severe pattern was observed in the Northwest, West and Southwest regions where peak monthly cases did not exceed 2500 (Fig S5). Some DENV cases during this period were serotyped as part of ongoing surveillance; the proportion of serotyped samples increased substantially in 2017 and 2018 when yearly number of cases were at their lowest (Fig S6). Most samples during this time were identified as DENV-1 or DENV-2, with only a small proportion DENV-3 or multi-serotype co-infections (Fig 3). When integrating previously reported data for the proportions of DENV serotypes from 2000 to 2013 *(71)*, it becomes clear that DENV-1 was practically absent from the country until it overtook all other serotypes between 2004 and 2007, and has since been co-circulating predominantly with DENV-2 (Fig 3A).

A notable anomaly was observed for DENV-1, as a large peak in case numbers was observed during 2017 followed by a decrease in 2018, when equal numbers for both

serotypes were once again observed (Fig 3B). This anomaly could be explained by the combination of low case numbers (Fig S6) and high serotyping output from the Center-north region (Fig S6). While the proportion of samples that were serotyped during 2017 increased across the country (Fig S7), the Center-north region serotyped ~95% of all reported cases while at the same time reporting the highest number of cases in the country (Fig S6, S7). In fact, the majority of the samples identified as DENV-1 during 2017 come from the Center-north (Fig S8). A breakdown of these patterns shows some common trends across regions (Fig 3C). Samples not identified as either DENV-1 or DENV-2 comprise a minority of the total serotyped samples, they were most prominently observed in the East and Northeast in 2016 but are completely absent from the north-western regions. The proportion of DENV-2 samples in relation to DENV-1 decreased between 2013 and 2016 in most of the country, except for the Southeast region. DENV-2 then increased in frequency between 2016 and 2017 in the Northeast and Southwest regions and began to increase in frequency in the rest of the country the following year (particularly in the West and Northwest; Fig 3C).

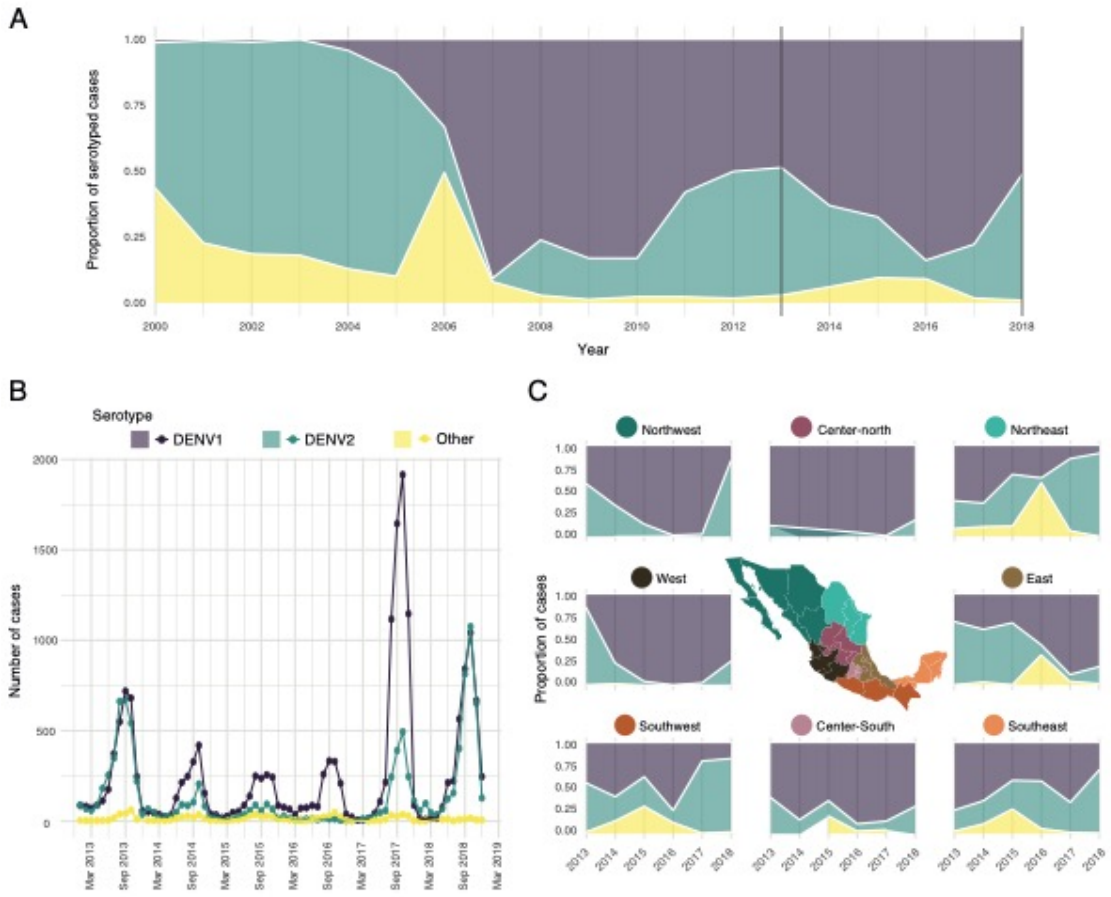


Figure 3. Epidemiological trends in DENV serotype frequencies in Mexico from 2013 to 2019. ***(A)*** Proportion of samples assigned to DENV serotypes between 2000 and 2018 in Mexico. Proportions are presented with reference to the total number of serotyped samples per year for all the country. Data reported in this study that shown between grey lines (2013 to 2018), while data from previous years was obtained from published sources *(71)*. DENV-3, DENV-4 and potential co-infections (reported between 2013 and 2018) are grouped into the 'Other' category. ***(B)*** Monthly number of samples assigned to each DENV serotype in Mexico from 2013 to 2018. ***(C)*** Breakdown of serotype proportions in different geographic regions in Mexico between 2013 and 2019, relative to the total number of serotyped samples in each location.

Phylogeography of DENV-1 and DENV-2 in Central America and Mexico

DENV-1 and DENV-2 in the Americas show some similarities in their phylogeographic dispersion patterns. ML trees for each DENV serotype in the Americas show that sequences from Mexico and Nicaragua cluster into a single lineage together

with sparse sequences from other Central American, North American or Caribbean nations (Fig S9), analogous to the CCNA lineage described above for CHIKV, but more geographically constrained (henceforth referred to as the Central America and Mexico lineages, CAM). Bayesian Skyline plots from these CAM lineages, which date back to the early 2000s for DENV-1 (median $MRCA_{DENV\text{-}1\text{-}CAM}$ = 2000.51, 95% HPD = 1997.88 – 2002.27) and to the late 1990s for DENV-2 (median $MRCA_{DENV\text{-}2\text{-}CAM}$ = 1996.82, 95% HPD = 1995.62 – 1997.79), show how the frequency of both viruses over time in Mexico match their periods of growth and decay adequately, particularly for the expansion of DENV-2 after 2000, the increase in frequency of DENV-1 around 2006 and the partial replacement of DENV-2 by DENV-1 following 2012 (Fig S10). This suggests that it is likely that the CCNA clade has governed the epidemiological trends of DENV-1 and DENV-2 in Mexico over the past 20 years. Discrete phylogeographic reconstruction indicates that the DENV-1 and DENV-2 CAM lineages predominantly circulate in Nicaragua and are periodically introduced into Mexico; we identified five different DENV-1 introductions and three DENV-2 introductions. The sampling timespan for both is similar; DENV-1 sequences were collected between 2004 and 2017 while DENV-2 sequences were collected between 2000 and 2019. The MRCAs for the different Mexico DENV sequence clusters show that these viruses have been circulating in the country since the late 1990s. Two DENV-2 clusters date back to 1999 (median $MRCA_{DENV\text{-}2\text{-}cluster\_1}$ = 1999.51, 95% HPD = 1998.13 – 2000.78; median $MRCA_{DENV\text{-}2\text{-}cluster\_2}$ = 1998.97, 95% HPD = 1997.35 – 2000.61) and circulated until 2013. Three DENV-1 Mexico clusters date back to the early 2000s (median $MRCA_{DENV\text{-}1\text{-}cluster\_1}$ = 2002.81, 95% HPD = 2001.62 – 2003.91; median $MRCA_{DENV\text{-}1\text{-}cluster\_2}$ = 2003.49, 95%

HPD = 2002.31 – 2004.18; median $MRCA_{DENV\text{-}1\text{-}cluster\_3}$ = 2004.03, 95% HPD = 2003.08 – 2004.80) and also circulated in the country at least until 2013. Further, a DENV-2 introduction is associated with a cluster dating back to 2008 (median $MRCA_{DENV\text{-}2\text{-}cluster\_3}$ = 2008.02, 95% HPD = 2007.15 – 2008.90) and a subsequent DENV-1 introduction is dated to 2009 (median $MRCA_{DENV\text{-}1\text{-}cluster\_4}$ = 2009.35, 95% HPD = 2008.28 – 2010.20; Fig S9).

The geographical sampling range of the DENV-1 genomic sequences is mostly limited to the Southeast region, whereas DENV-2 sequences are available for a broader range of locations across southern and central Mexico. A discrete phylogeographic reconstruction shows that the three DENV-2 clusters within the country originated in the Southwest (Fig 4), from where they spread northward via the East and West regions and into the Yucatán peninsula. A BSSVS analysis suggests that movements between Nicaragua and the Southwest region of Mexico explain the international importations of DENV-2 (BF = 1124.01, PP = 0.99), from where the virus is exported into other regions in the country. Significant movements from the Southwest into the West (BF = 2929.53, PP > 0.99), Southeast (BF = 1270.88, PP > 0.99) and East (BF = 56.08, PP = 0.90) regions of Mexico are inferred, as well as movements from the Southwest across the southern border into Guatemala (BF = 134.65, PP = 0.96). Movements from the West to the Center-south regions of Mexico (namely the state of Morelos) are also well supported (BF = 83.25, PP = 0.93). As with the CHIKV dataset, the large number of sequences from Nicaragua and scarcity of sequences from other countries in the region likely affect the BSSVS estimates and phylogeographic reconstructions (Table S3).

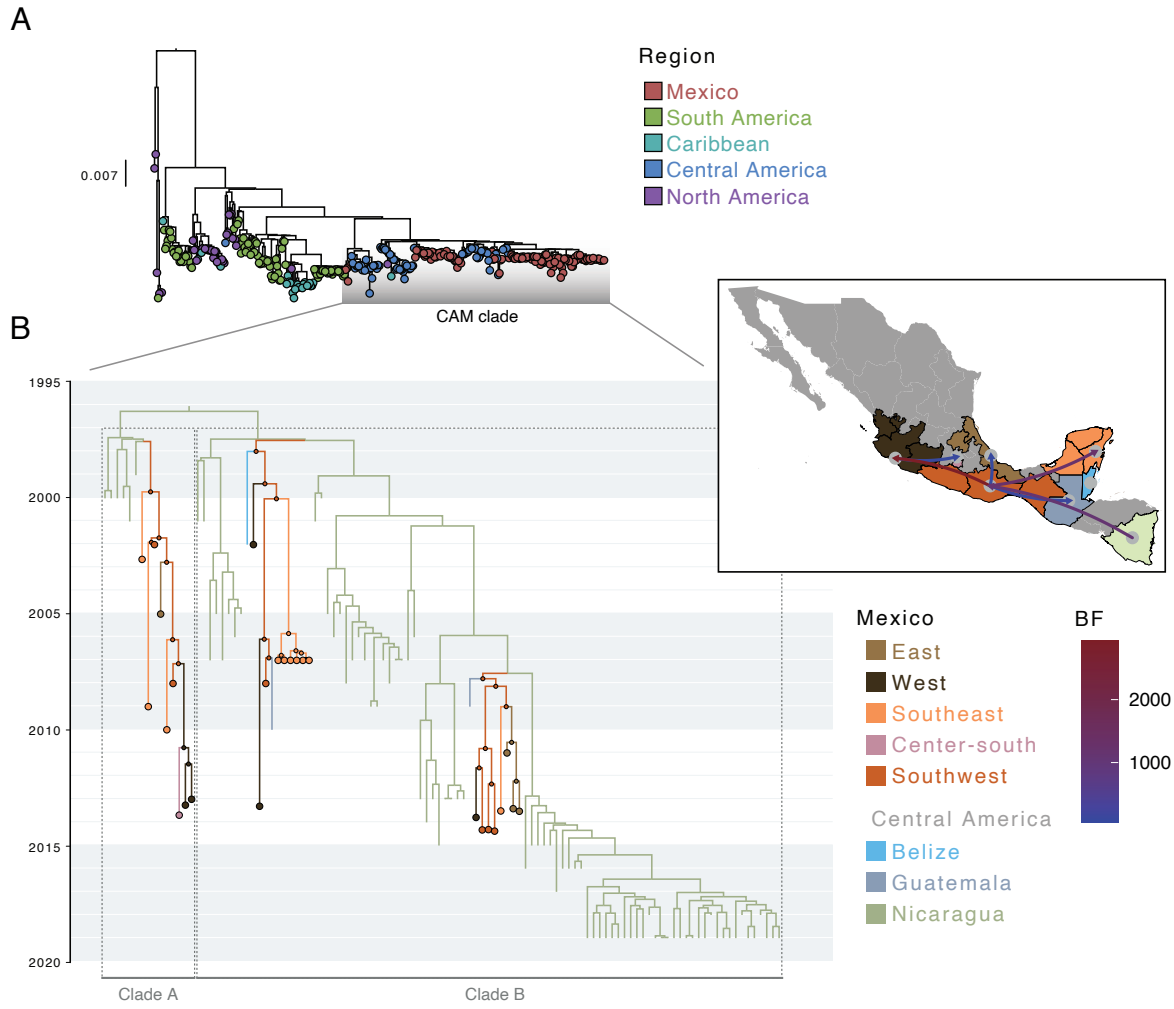


Figure 4. Importation and domestic spread of DENV-2 in Mexico. ***(A)*** ML phylogenetic trees of all DENV-2 complete genome sequences from the Americas included in our analysis. Tree tips are coloured by region where they were collected. ***(B)*** DTA phylogeographic analysis of the DENV-2 CAM clade by sampling location in Mexico. Tip symbols and nodes are included for locations within Mexico and are coloured by region. Distinct clades are marked as Clade A (which includes one Mexico cluster) and Clade B (which includes two Mexico clusters). The map in the inset shows pairs of locations where transition rates significantly explain the phylogenetic diffusion process, as inferred by a BSSVS analysis. Transition rates with a posterior probability > 0.5 are shown and coloured by Bayes Factor (BF).

Human mobility trends in Mexico and neighbouring countries

The spread of viruses like DENV can be affected by anthropogenic factors such as human mobility across geographic regions *(6)*. While travel volume metrics such as flight data can provide useful estimates for connectivity between countries, these values will not represent sources of land-based mobility which could account for an important proportion of human movements in regions like Central America and Mexico *(102, 103)*, especially given the potential role of Mexico as a transit location *(104)* between countries in Central America and the USA. Therefore we used cell-phone usage data collected between October 2019 and late February 2020 to generate quantitative connectivity estimates between geographic regions in Mexico and the Central American countries south of the border (Belize, Guatemala, El Salvador, Honduras and Nicaragua; as described in *(94)*). Data in Mexico was collected for movements between 2463 individual municipality-level locations, and for movements between these locations and the aforementioned countries over a period of 18 weeks between 2019-10-27 and 2020-02-29. The connectivity between locations (represented as the relative flow, f_r) directly quantifies the total number of individual movements between pairs of locations in different weeks of late 2019 and early 2020, and normalises these anonymised movements relative to the largest number of movements between location pairs over that time period *(94)*.

Aggregated median f_r estimates aggregated over the complete time period between October 2019 and February 2020 capture high flow between states along the eastern Gulf coast of Mexico (particularly the Southeast region) and states in central and

northern regions. Notably, the state of Quintana Roo in the Southeast shows high connectivity with states across northern Mexico but limited human flow with the Southwest. In fact, the only southwestern state with high connectivity to other states in the country is Oaxaca which received high flow frequencies from the Center-north over the entire data collection period, and from the West during late December and early January. Human flow between the Southwest and West regions along the Pacific coast appears limited, as the latter region was more highly connected to the Southeast, especially during early November 2019 and late February 2020 (Fig 5). Movements across international borders became more frequent during the last week of 2019 and first week of 2020. Persistent movements between El Salvador and both Guatemala and Honduras are visible across time, and human flows into El Salvador from Nicaragua and into Guatemala from Belize increased during the end of 2019. Another important international border lies between Quintana Roo in the Southeast and Belize, where high median f_r estimates are recorded throughout the data collection period. Human movements from Guatemala into the states in the Southwest of Mexico were limited (Fig 5).

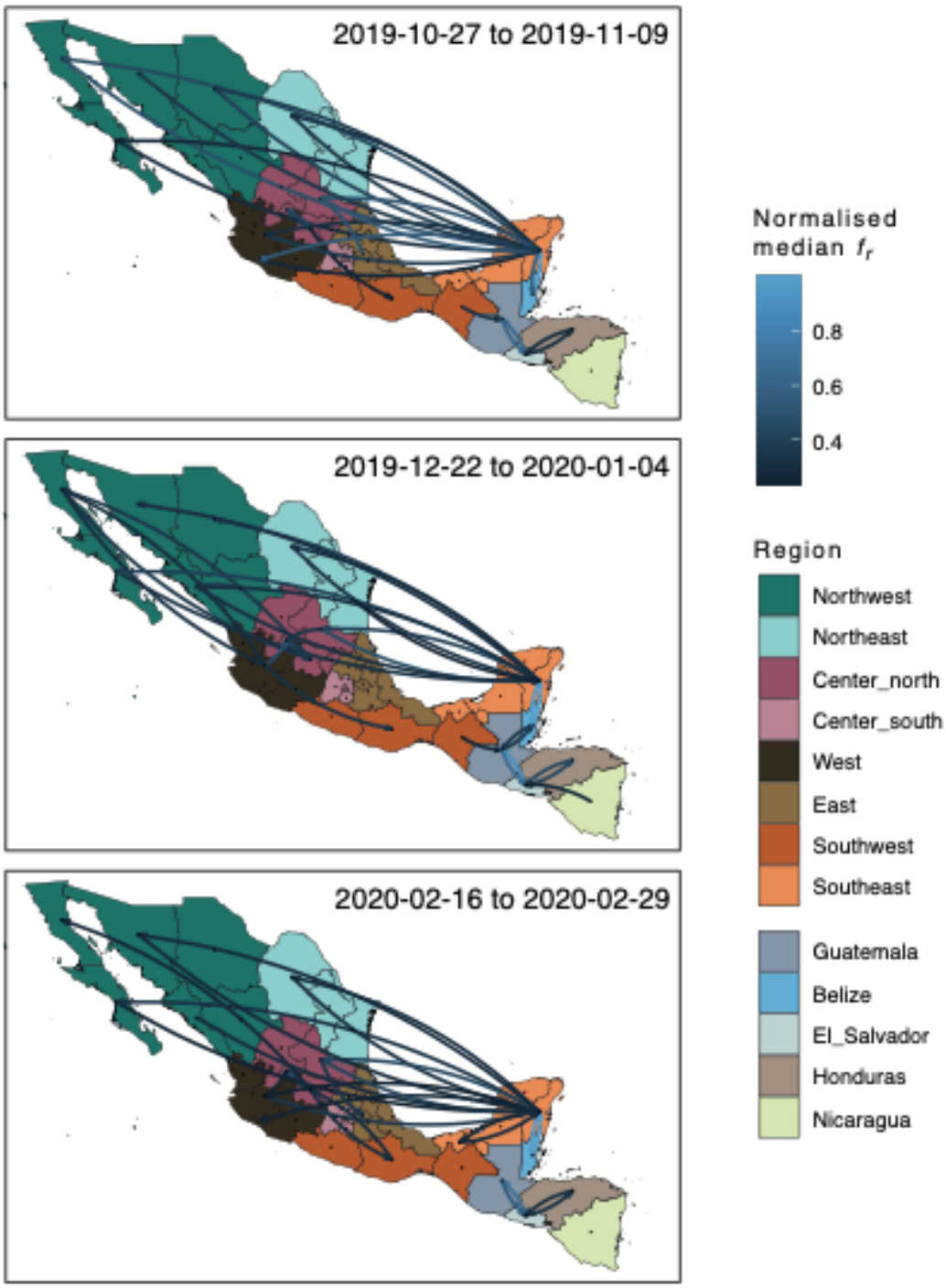


Figure 5. Human mobility patterns in Mexico and Central America between October 2019 and March 2020. Map of Mexico (coloured by geographic regions) and the Central American countries of Belize, Guatemala, El Salvador, Honduras and Nicaragua shows the higher 90th percentile of all median relative flow estimates (f_r) obtained from anonymised smartphone geolocation mapping. Weekly movements between municipalities in Mexico between 2019-10-27 and 2020-02-29 were

aggregated by region, and median relative human flow over bi-weekly time periods were estimated. Figure shows a selection of three bi-weekly period at the beginning of the data collection period, during the end of 2019 and start of 2020 and the end of the data collection period. Arrows on map are coloured by the bi-weekly median f_r normalised by the maximum median f_r for the entire data collection period.

Discussion

The transmission of *Aedes*-borne viruses such as DENV, CHIKV and ZIKV depends on numerous environmental, entomological and anthropogenic factors that affect the suitability and distribution of their vector species *Ae. aegypti* and *Ae. albopictus (105, 106)*. Comparison of the dissemination of arboviruses that share one or more vector species represents a useful tool to assess the similarities and differences between their transmission drivers across regions and timescales. However, multiple factors play a role in how pathogens are transmitted and there are important distinctions between transient epidemics (driven by viral importation; transmission for a limited time) and the establishment of a virus in a region (persistent or seasonal transmission over longer time periods). Human mobility represents a key determinant of the spread of viral pathogens *(107)*, including arboviruses such as DENV *(6, 77)*. Whilst human mobility and behaviour will partially drive the spatial spread of arboviruses and occurrence of individual epidemics, the ability of a virus to establish sustained or seasonal transmission in a given recipient geographical area will likely depend on its ability to persistently infect the available local vector populations and immunity within the local host population.

While DENV is known to affect states in the centre and south of Mexico, its circulation in the country is likely to be affected by the incidence of DENV infection in neighbouring countries. National borders offer no resistance to the natural movement of vector species *(105)* and cross-border human mobility is prevalent in this region, particularly in the context of international migration from Central America into the USA through Mexico *(104)*. We find that periodic DENV importations occurred through the 2000s and 2010s, and that these could partially explain the shifting dominance of specific serotypes through time in Mexico. DENV clade replacements have been described widely in Asia *(65, 66, 77)* and South America *(108)*, and our results suggest that a similar phenomenon could take place in Central America.

Our phylogeographic analysis of DENV-2 indicates multiple introductions to Mexico from Nicaragua, but the source of these inferred introductions is likely biased by sampling differences in the region: while Nicaragua features a considerable genomic representation of the virus as a result of longitudinal paediatric cohort studies *(59, 82)*, which likely capture a greater proportion of all cases in the country given that DENV has high attack rates within these younger age groups (*56, 59*), other countries like Belize, Guatemala, Honduras and El Salvador are poorly represented in our genomic data. Given the geography of the region, it is likely that some of these countries play an important role in the regional spatial dynamics of the virus. This is further evidenced by the phylogenetic placement of two DENV-2 sequences, one from Belize and one from Guatemala, at the base of two different Mexico clusters. The most probable location for the MRCA of each of the Mexico DENV-2 clusters was inferred to be the Southwest

region, where Chiapas is located, a state that shares a border with Guatemala. The Southeast region, which encompasses the Yucatán peninsula, is also well represented in our genomic sample and shares an international border with Belize, but we find that sequences from the Southeast are introduced from the Southwest of Mexico rather than from other Central American countries (Fig 4). Further, sequences from both central coastal regions, East and West, appear to have originated from the Southwest. These patterns are similar to those observed for CHIKV, where four out of six CHIKV transmission clusters in Mexico have ancestral location in the Southwest (even when the geographic sampling distribution of CHIKV in the country is much broader than that of DENV-2), and to previous analyses for ZIKV *(97)*. Thézé et al. showed that ZIKV was introduced into the state of Chiapas probably from Honduras, from where it spread into the states of Oaxaca and Guerrero across the Pacific coast *(97)*. The repeated appearance of the Southwest region as an ancestral location within Mexico of transmission clusters for DENV, CHIKV and ZIKV provides strong evidence that this region tends to act as a port of entry for seasonal and emerging *Aedes*-borne arboviruses into the country.

The observation of such persistent dispersal trends suggests that the importation and spread of CHIKV and ZIKV in Mexico during 2015 and 2016 were likely driven by similar factors. Furthermore, our results suggest that the serotype replacement dynamics of DENV during the 1990s and 2000s might also be influenced by some of these factors. Viral establishment and seasonality, as is the case with DENV, are determined by previous exposure and protective immunity, immunological waning and the replenishment of susceptible hosts *(1, 26, 27, 65, 67)*. However, the patterns that we

show here suggest that arboviral spatial invasion can present similarities between viruses regardless of the presence or absence of protective immunity from previous exposure. It is nonetheless intriguing that the connectivity between regions, as inferred from smartphone usage, shows that human mobility between the Southwest and the rest of Mexico is limited (Fig 5). This could be explained by the role that other factors play in the transmission of arboviruses (such as environmental conditions and vector suitability) or limited smartphone usage in this region of the country. Similarly, the high frequency of human mobility between Belize and the Southeast suggests that the Yucatan peninsula could display distinct epidemiological trends for arboviruses as is visible from the late CHIKV peak in 2015 compared to other regions in the country (Fig 1C). Greater genomic sampling in the countries neighbouring Mexico's southern states and a more detailed exploration of arboviral epidemiological trends in these locations would expand on the local dynamics and spread of these viruses in the region. Also, the lack of smartphone usage data that overlaps with the genomic sampling time period limits our capacity to formally test the effect of human mobility in the spread of all three viruses through a phylogenetic GLM (*113*). This highlights the importance of utilising multiple data streams to maximise the inferential power derived from genomic data.

These predictable spatial patterns could play an important role in the design of surveillance programs and public health interventions, and also on the implementation of vaccine trials and programmes. Given that previous immunity is an important covariate when interpreting results for novel vaccine trials, the higher probability of new arboviral importations in the Southwest states make this region more likely to account

for areas with high seroprevalence, as has been proposed previously *(13)*. Other Mexican regions from where domestic importations from the Southwest occur, such as the West and Southeast, have also been previously shown to present high DENV seroprevalence *(109, 110)*. Conversely, the higher probability of new importations is also an important factor that increases the availability of susceptible individuals, as sufficiently distinct genetic variants (as would be the case for a new or low prevalence DENV serotype) are more likely to be neutralised by the pre-existing population immunity. This can be an important consideration particularly for Phase II and Phase III vaccine trial stages when vaccine efficacy is a key measured outcome, as the reduction of new cases can delay the data collection phase *(111)*.

While our analyses show similar patterns across viruses and timescales, it's important to account for sampling bias in the interpretation of these results. Arboviral disease samples are naturally more likely to come from the southern regions where the prevalence is highest *(72)*, and this is reflected in the representation of sequences for all three viruses in our data sets. The limited number of genomes from northern Mexico limits our capacity to quantify the role that other regions could play in the broader spatial dynamics of arboviruses in the country. This might be particularly relevant for analysing exchanges between Mexico and the southern United States, where these pathogens have been detected and where substantial mobility across their shared border could provide ample opportunities for viral importations. Incomplete sampling of the currently identified clusters also introduces uncertainty to our estimates for the introduction dates. For transmission clusters preceded by long branches in the time-calibrated phylogenies,

the introduction events themselves could have occurred at a time between the cluster's *t*MRCA and the age of the preceding node. Long branches on our time-calibrated trees may be an indicator of unsampled viral genetic diversity that can also bias inferred locations of the cluster's ancestral nodes *(112)*. Nonetheless, the historical DENV epidemiological patterns and the unfolding of the ZIKV and CHIKV epidemics suggest are consistent with our finding that the southern regions are indeed important hubs of arboviral importations and transmission. It is therefore likely that our conclusions are sound, and that multiple arboviruses tend to be repeatedly imported and established in the Southwestern regions of Mexico from where they seed epidemics in other regions *(97)*.

Multiple tools have been developed to guide the implementation of interventions to control the transmission of dengue fevers and other arbovirus diseases *(72, 105, 106)*. Our results add to the current knowledge of arbovirus transmission in Mexico and highlight the occurrence of potentially predictable spatial invasion dynamics. Incorporating this data into current epidemiological disease models may help to enhance the accuracy of current risk maps for new emerging arboviruses in Mexico and Central America. Finally, the key role of Mexico's southern border in the spread of arboviruses prompts the necessity to better understand the role of anthropogenic factors in the transmission dynamics of these pathogens, particularly as it concerns the effects of land-based migration and environmental suitability for vector species.

References

1. M. Oki, T. Yamamoto, Climate Change, Population Immunity, and Hyperendemicity in the Transmission Threshold of Dengue, *PLoS One* **7** (2012), doi:10.1371/journal.pone.0048258.
2. D. W. Redding, R. Gibb, C. C. Dan-Nwafor, E. A. Ilori, R. U. Yashe, S. H. Oladele, M. O. Amedu, A. Iniobong, L. A. Attfield, C. A. Donnelly, I. Abubakar, K. E. Jones, C. Ihekweazu, Geographical drivers and climate-linked dynamics of Lassa fever in Nigeria, *Nat. Commun.* **12**, 1–10 (2021).
3. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, MERS-CoV spillover at the camel-human interface, *Elife* **7**, 1–23 (2018).
4. A. M. Kilpatrick, S. E. Randolph, Drivers, dynamics, and control of emerging vector-borne zoonotic diseases, *Lancet* **380**, 1946–1955 (2012).
5. N. R. Faria, M. U. G. Kraemer, S. C. Hill, J. Goes de Jesus, R. S. de Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C. H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolai, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa, S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. dos Santos, C. D. dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira, R. G. Abreu, M. H. O. Garcia, R. V. Alves, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, O. G. Pybus, Genomic and epidemiological monitoring of yellow fever virus transmission potential, *bioRxiv* **899**, 894–899 (2018).
6. O. M. Allicock, N. Sahadeo, P. Lemey, A. J. Auguste, M. A. Suchard, A. Rambaut, C. V. F. Carrington, Determinants of dengue virus dispersal in the americas, *Virus Evol.* **6**, 1–13 (2020).
7. P. Reiter, S. Lathrop, M. Bunning, B. Biggerstaff, D. Singer, T. Tiwari, L. Baber, M. Amador, J. Thirion, J. Hayes, C. Seca, J. Mendez, B. Ramirez, J. Robinson, J. Rawlings, V. Vorndam, S. Waterman, D. Gubler, G. Clark, E. Hayes, Texas lifestyle limits transmission of dengue virus, *Emerg. Infect. Dis.* **9**, 86–89 (2003).
8. N. R. Faria, R. Do Socorro Da Silva Azevedo, M. U. G. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Thézé, M. B. Bonsall, T. A. Bowden, I. Rissanen, I. M. Rocco, J. S. Nogueira, A. Y. Maeda, F. G. Da Silva Vasami, F. L. De Lima Macedo, A. Suzuki, S. G. Rodrigues, A. C. R. Cruz, B. T. Nunes, D. B. De Almeida Medeiros, D. S. G. Rodrigues, A. L. N. Queiroz, E. V. P. Da Silva, D. F. Henriques, E. S. T. Da Rosa, C. S. De Oliveira, L. C. Martins, H. B. Vasconcelos, L. M. N. Casseb, D. De Brito Simith, J. P. Messina, L. Abade, J. Lourenço, L. C. Alcantara, M. M. De Lima, M. Giovanetti, S. I. Hay, R. S. De Oliveira, P. Da Silva Lemos, L. F. De Oliveira, C. P. S. De Lima, S. P. Da Silva, J. M. De Vasconcelos, L. Franco, J. F. Cardoso, J. L. Da Silva Gonçalves Vianez-Júnior, D. Mir, G. Bello, E. Delatorre, K. Khan, M. Creatore,

G. E. Coelho, W. K. De Oliveira, R. Tesh, O. G. Pybus, M. R. T. Nunes, P. F. C. Vasconcelos, Zika virus in the Americas: Early epidemiological and genetic findings, *Science (80-.).* **352**, 345–349 (2016).
9. M. U. G. Kraemer, C. H. Yang, B. Gutierrez, C. H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino, The effect of human mobility and control measures on the COVID-19 epidemic in China, *Science (80-.).* **368**, 493–497 (2020).
10. M. Churakov, C. J. Villabona-Arenas, M. U. G. Kraemer, H. Salje, S. Cauchemez, Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony, *PLoS Negl. Trop. Dis.* **13**, 1–13 (2019).
11. N. Ferguson, R. Anderson, S. Gupta, The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 790–794 (1999).
12. H. Salje, J. Lessler, I. M. Berry, M. C. Melendrez, T. Endy, S. Kalayanarooj, A. A-nuegoonpipat, S. Chanama, S. Sangkijporn, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size, *Science (80-.).* **355**, 1302–1306 (2017).
13. I. Rodríguez-Barraquer, H. Salje, D. A. T. Cummings, Opportunities for improved surveillance and control of infectious diseases from age-specific case data, *Elife* **8**, e45474 (2019).
14. M. A. Robert, R. C. Christofferson, P. D. Weber, H. J. Wearing, Temperature impacts on dengue emergence in the United States: Investigating the role of seasonality and climate change, *Epidemics* **28**, 100344 (2019).
15. A. R. Akhmetzhanov, Y. Asai, H. Nishiura, Quantifying the seasonal drivers of transmission for Lassa fever in Nigeria, *Philos. Trans. R. Soc. B Biol. Sci.* **374** (2019), doi:10.1098/rstb.2018.0268.
16. D. B. George, C. T. Webb, M. L. Farnsworth, T. J. O'Shea, R. A. Bowen, D. L. Smith, T. R. Stanley, L. E. Ellison, C. E. Rupprecht, Host and viral ecology determine bat rabies seasonality and maintenance, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10208–10213 (2011).
17. Y. Xiao, J. C. Beier, R. S. Cantrell, C. Cosner, D. L. DeAngelis, S. Ruan, Modelling the Effects of Seasonality and Socioeconomic Impact on the Transmission of Rift Valley Fever Virus, *PLoS Negl. Trop. Dis.* **9** (2015), doi:10.1371/journal.pntd.0003388.
18. D. J. Gubler, N. Vasilakis, D. Musso, History and Emergence of Zika Virus, *J. Infect. Dis.* **216**, S860–S867 (2017).
19. M. K. Njenga, L. Nderitu, J. P. Ledermann, A. Ndirangu, C. H. Logue, C. H. L. Kelly, R. Sang, K. Sergon, R. Breiman, A. M. Powers, Tracking epidemic Chikungunya virus into the Indian Ocean from East Africa, *J. Gen. Virol.* **89**, 2754–2760 (2008).
20. T. E. Morris, Reemergence of Chikungunya virus, *J. Virol.* **88**, 11644–11647 (2014).

21. J. J. Muyembe-Tamfum, S. Mulangu, J. Masumu, J. M. Kayembe, A. Kemp, J. T. Paweska, Ebola virus outbreaks in Africa: Past and present, *Onderstepoort J. Vet. Res.* **79**, 1–8 (2012).
22. P. Daszak, P. K. Plowright, J. H. Epstein, J. Pulliam, S. Abdul Rahman, H. E. Field, A. A. Bin Jamaluddin, S. H. Sharifah, C. S. Smith, K. J. Olival, S. Luby, K. Halpin, A. D. Hyatt, A. A. Cunningham, in *Disease Ecology: Community Structure and Pathogen Dynamics*, (2007), pp. 186–201.
23. L. B. Carrington, C. P. Simmons, Human to mosquito transmission of dengue viruses, *Front. Immunol.* **5**, 1–8 (2014).
24. P. S. J. Wong, M. zhi I. Li, C. S. Chong, L. C. Ng, C. H. Tan, Aedes (Stegomyia) albopictus (Skuse): A Potential Vector of Zika Virus in Singapore, *PLoS Negl. Trop. Dis.* **7**, 1–5 (2013).
25. I. M. Mombo, D. Jiolle, D. Fontenille, C. Paupy, E. M. Leroy, Zika Virus in Gabon (Central Africa) – 2007 : A New Threat from Aedes albopictus ?, **8**, 1–6 (2014).
26. R. K. Borchering, A. T. Huang, L. Mier-y-Teran-Romero, D. P. Rojas, I. Rodriguez-Barraquer, L. C. Katzelnick, S. D. Martinez, G. D. King, S. C. Cinkovich, J. Lessler, D. A. T. Cummings, Impacts of Zika emergence in Latin America on endemic dengue transmission, *Nat. Commun.* **10** (2019), doi:10.1038/s41467-019-13628-x.
27. G. S. Ribeiro, M. Kikuti, L. B. Tauro, L. C. J. Nascimento, C. W. Cardoso, G. S. Campos, A. I. Ko, S. C. Weaver, M. G. Reis, U. Kitron, I. A. D. Paploski, M. M. O. Silva, A. M. Kasper, A. S. Tavares, J. S. Cruz, P. S. S. Moreira, R. O. Anjos, J. M. G. Araújo, R. Khouri, S. I. Sardi, Does immunity after Zika virus infection cross-protect against dengue?, *Lancet Glob. Heal.* **6**, e140–e141 (2018).
28. V. A. Mugabe, L. S. Borja, C. W. Cardoso, S. C. Weaver, M. G. Reis, U. Kitron, G. S. Ribeiro, Changes in the dynamics of dengue incidence in South and Central America are possibly due to cross-population immunity after Zika virus epidemics, *Trop. Med. Int. Heal.* **26**, 272–280 (2021).
29. D. E. Carey, Chikungunya and dengue: A case of mistaken identity?, *J. Hist. Med. Allied Sci.* **26**, 243–262 (1971).
30. E. Gould, J. Pettersson, S. Higgs, R. Charrel, X. de Lamballerie, Emerging arboviruses: Why today?, *One Heal.* **4**, 1–13 (2017).
31. S. C. Weaver, Arrival of Chikungunya Virus in the New World: Prospects for Spread and Impact on Public Health, *PLoS Negl. Trop. Dis.* **8**, 6–9 (2014).
32. R. S. Lanciotti, A. M. Valadere, Transcontinental movement of Asian Genotype Chikungunya virus, *Emerg. Infect. Dis.* **20**, 1400–1402 (2014).
33. R. S. Nasci, Movement of Chikungunya virus into the Western Hemisphere, *Emerg. Infect. Dis.* **20**, 1394–1395 (2014).
34. M. S. Cunha, P. A. G. Costa, I. A. Correa, M. R. M. de Souza, P. T. Calil, G. P. D. da Silva, S. M. Costa, V. W. P. Fonseca, L. J. da Costa, Chikungunya Virus: An Emergent Arbovirus to the South American Continent and a Continuous Threat to the World, *Front. Microbiol.* **11** (2020), doi:10.3389/fmicb.2020.01297.
35. I. M. Berry, W. Rutvisuttinunt, R. Sippy, E. Beltran-ayala, K. Figueroa, S. Ryan, A. Srikanth, A. M. Stewart-ibarra, T. Endy, R. G. Jarman, The origins of dengue and

chikungunya viruses in Ecuador following increased migration from Venezuela and Colombia, **8**, 1–12 (2020).
36. D. Camacho, J. Reyes, A. Negredo, L. Hernández, M. Sánchez-Seco, G. Comach, Asian genotype of Chikungunya virus circulating in Venezuela during 2014, *Acta Trop.* **174**, 88–90 (2017).
37. I. Leparc-Goffart, A. Nougairede, S. Cassadou, C. Prat, X. De Lamballerie, Chikungunya in the Americas, *Lancet* **383**, 514 (2014).
38. M. R. T. Nunes, N. R. Faria, J. M. de Vasconcelos, N. Golding, M. U. G. Kraemer, L. F. de Oliveira, R. do S. da Silva Azevedo, D. E. A. da Silva, E. V. P. da Silva, S. P. da Silva, V. L. Carvalho, G. E. Coelho, A. C. R. Cruz, S. G. Rodrigues, J. L. da Silva Gonçalves Vianez, B. T. D. Nunes, J. F. Cardoso, R. B. Tesh, S. I. Hay, O. G. Pybus, P. F. da Costa Vasconcelos, Emergence and potential for spread of Chikungunya virus in Brazil, *BMC Med.* **13** (2015), doi:10.1186/s12916-015-0348-x.
39. F. G. Naveca, I. Claro, M. Giovanetti, J. G. de Jesus, J. Xavier, F. C. de M. Iani, V. A. do Nascimento, V. C. de Souza, P. P. Silveira, J. Lourenço, M. Santillana, M. U. G. Kraemer, J. Quick, S. C. Hill, J. Thézé, R. D. de O. Carvalho, V. Azevedo, F. C. da S. Salles, M. R. T. Nunes, P. da S. Lemos, D. da S. Candido, G. de C. Pereira, M. A. A. Oliveira, C. A. R. Meneses, R. M. Maito, C. R. S. B. Cunha, D. P. de S. Campos, M. da C. Castilho, T. C. da S. Siqueira, T. M. Terra, C. F. C. de Albuquerque, L. N. da Cruz, A. L. de Abreu, D. V. Martins, D. S. de M. V. Simoes, R. S. de Aguiar, S. L. B. Luz, N. Loman, O. G. Pybus, E. C. Sabino, O. Okumoto, L. C. J. Alcantara, N. R. Faria, Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon, *PLoS Negl. Trop. Dis.* **13**, 1–21 (2018).
40. J. Xavier, M. Giovanetti, V. Fonseca, J. Thézé, T. Gräf, A. Fabri, J. G. de Jesus, M. C. L. de Mendonça, C. D. dos Santos Rodrigues, M. A. Mares-Guia, C. C. dos Santos, S. F. de Oliveira Tosta, D. Candido, R. M. R. Nogueira, A. L. de Abreu, W. K. Oliveira, C. F. Campelo de Albuquerque, A. Chieppe, T. de Oliveira, P. Brasil, G. Calvet, P. C. Sequeira, N. R. Faria, A. M. B. de Filippis, L. C. J. Alcantara, Circulation of chikungunya virus East/Central/ South African lineage in Rio de Janeiro, Brazil, *PLoS One* **14**, 1–14 (2019).
41. E. C. de Oliveira, V. Fonseca, J. Xavier, T. Adelino, I. M. Claro, A. Fabri, E. M. Macario, A. E. Viniski, C. L. C. Souza, E. S. G. da Costa, C. S. de Sousa, F. G. D. Duarte, A. C. de Medeiros, C. F. C. de Albuquerque, R. V. Cunha, N. F. O. De Moura, A. M. B. de Filippis, T. de Oliveira, J. Lourenço, A. L. de Abreu, L. C. J. Alcantara, M. Giovanetti, Short report: Introduction of chikungunya virus ecsa genotype into the brazilian midwest and its dispersion through the americas, *PLoS Negl. Trop. Dis.* **15**, 1–10 (2021).
42. L. Mowatt, S. T. Jackson, Chikungunya in the Caribbean: An Epidemic in the Making, *Infect. Dis. Ther.* **3**, 63–68 (2014).
43. M. A. Johansson, Chikungunya on the move, *Trends Parasitol.* **31**, 43–45 (2015).
44. M. A. Johansson, A. M. Powers, N. Pesik, N. J. Cohen, J. E. Staples, Nowcasting the Spread of Chikungunya Virus in the Americas, **9** (2014), doi:10.1371/journal.pone.0104915.
45. K. Khan, I. Bogoch, J. S. Brownstein, J. Miniota, A. Nicolucci, W. Hu, E. O. Nsoesie, M. Cetron, M. I. Creatore, M. German, A. Wilder-Smith, Assessing the

Origin of and Potential for International Spread of Chikungunya Virus from the Caribbean, *PLoS Curr.* , 4055609 (2014).
46. L. E. Escobar, H. Qiao, A. T. Peterson, Forecasting Chikungunya spread in the Americas via data-driven empirical approaches, *Parasites and Vectors* **9**, 1–12 (2016).
47. M. Nava-Frías, R. E. Searcy-Pavía, C. A. Juárez-Contreras, A. Valencia-Bautista, Chikungunya fever: Current status in Mexico, *Bol. Med. Hosp. Infant. Mex.* **73**, 67–74 (2016).
48. D. Nunez-Avellaneda, C. Tangudu, J. Barrios-Palacios, M. I. Salazar, C. Machain-Williams, J. Cisneros-Pano, L. A. McKeen, B. J. Blitvich, Chikungunya in Guerrero, Mexico, 2019 and Evidence of Gross Underreporting in the Region, *Am. J. Trop. Med. Hyg.* **105**, 1281–1284 (2021).
49. J. A. Díaz-Quiñonez, J. Ortiz-Alcántara, D. E. Fragoso-Fonseca, F. Garcés-Ayala, N. Escobar-Escamilla, M. Vázquez-Pichardo, A. Núñez-León, M. de la L. Torres-Rodríguez, B. Torres-Longoria, I. López-Martínez, C. Ruíz-Matus, P. Kuri-Morales, J. E. Ramírez-González, Complete genome sequences of chikungunya virus strains isolated in Mexico: First detection of imported and autochthonous cases, *Genome Announc.* **3** (2015), doi:10.1128/genomeA.00300-15.
50. S. V. Laredo-Tiscareño, C. MacHain-Williams, M. A. Rodríguez-Pérez, J. A. Garza-Hernandez, G. L. Doria-Cobos, R. C. Cetina-Trejo, L. A. Bacab-Cab, C. S. Tangudu, J. Charles, E. J. De Luna-Santillana, J. E. Garcia-Rejon, B. J. Blitvich, Arbovirus surveillance near the Mexico-U.S. Border: Isolation and sequence analysis of chikungunya virus from patients with dengue-like symptoms in reynosa, tamaulipas, *Am. J. Trop. Med. Hyg.* **99**, 191–194 (2018).
51. J. E. Muñoz-Medina, M. A. Garcia-Knight, A. Sanchez-Flores, I. E. Monroy-Muñoz, R. Grande, J. Esbjörnsson, C. E. Santacruz-Tinoco, C. R. González-Bonilla, Evolutionary analysis of the Chikungunya virus epidemic in Mexico reveals intra-host mutational hotspots in the E1 protein, *PLoS One* **13**, 1–18 (2018).
52. K. A. Galán-Huerta, E. Martínez-Landeros, J. L. Delgado-Gallegos, S. Caballero-Sosa, I. R. Malo-García, I. Fernández-Salas, J. Ramos-Jiménez, A. M. Rivas-Estilla, Molecular and clinical characterization of Chikungunya virus infections in Southeast Mexico, *Viruses* **10**, 1–18 (2018).
53. O. M. Allicock, P. Lemey, A. J. Tatem, O. G. Pybus, S. N. Bennett, B. A. Mueller, M. A. Suchard, J. E. Foster, A. Rambaut, C. V. F. Carrington, Phylogeography and population dynamics of dengue viruses in the Americas, *Mol. Biol. Evol.* **29**, 1533–1543 (2012).
54. D. J. Gubler, Dengue and Dengue Hemorrhagic Fever, *Clin. Microbiol. Rev.* **11**, 480–496 (1998).
55. F. Pinheiro, M. Nelson, Re-Emergence of Dengue and Emergence of Dengue Haemorrhagic Fever in the Americas, *Dengue Bull.* **21**, 1–6 (1997).
56. A. Gordon, G. Kuan, J. C. Mercado, L. Gresh, W. Avilés, A. Balmaseda, E. Harris, The Nicaraguan Pediatric Dengue Cohort Study: Incidence of Inapparent and Symptomatic Dengue Virus Infections, 2004-2010, *PLoS Negl. Trop. Dis.* **7**, 2004–2010 (2013).

57. M. F. Vincenti-Gonzalez, A. Tami, E. F. Lizarazo, M. E. Grillet, ENSO-driven climate variability promotes periodic major outbreaks of dengue in Venezuela, *Sci. Rep.* **8**, 1–11 (2018).
58. R. Sippy, D. Herrera, D. Gaus, R. E. Gangnon, J. A. Patz, J. E. Osorio, Seasonal patterns of dengue fever in rural Ecuador: 2009-2016, *PLoS Negl. Trop. Dis.* **13**, 2009–2016 (2019).
59. A. Balmaseda, K. Standish, J. C. Mercado, J. C. Matute, Y. Tellez, S. Saborío, S. N. Hammond, A. Nuñez, W. Avilés, M. R. Henn, E. C. Holmes, A. Gordon, J. Coloma, E. Harris, Trends in patterns of dengue transmission over four years of a pediatric cohort study in Nicaragua, *J Infect Dis.* **201**, 5–14 (2010).
60. R. N. Guo, J. Y. Lin, L. H. Li, C. W. Ke, J. F. He, H. J. Zhong, H. Q. Zhou, Z. Q. Peng, F. Yang, W. J. Liang, The prevalence and endemic nature of dengue infections in Guangdong, South China: An epidemiological, serological, and etiological study from 2005-2011, *PLoS One* **9** (2014), doi:10.1371/journal.pone.0085596.
61. I. Morales, H. Salje, S. Saha, E. S. Gurley, Seasonal distribution and climatic correlates of dengue disease in Dhaka, Bangladesh, *Am. J. Trop. Med. Hyg.* **94**, 1359–1361 (2016).
62. S. Polwiang, The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017), *BMC Infect. Dis.* **20**, 1–10 (2020).
63. S. B. Halstead, Dengue virus-mosquito interactions, *Annu. Rev. Entomol.* **53**, 273–291 (2008).
64. A. M. Stewart-Ibarra, R. Lowe, Climate and non-climate drivers of dengue epidemics in southern coastal Ecuador, *Am. J. Trop. Med. Hyg.* **88**, 971–981 (2013).
65. B. Adams, E. C. Holmes, C. Zhang, M. P. Mammen, S. Nimmannitya, S. Kalayanarooj, M. Boots, Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14234–14239 (2006).
66. C. Zhang, M. P. Mammen, P. Chinnawirotpisan, C. Klungthong, P. Rodpradit, P. Monkongdee, S. Nimmannitya, S. Kalayanarooj, E. C. Holmes, Clade Replacements in Dengue Virus Serotypes 1 and 3 Are Associated with Changing Serotype Prevalence, *J. Virol.* **79**, 15123–15130 (2005).
67. M. Recker, K. B. Blyuss, C. P. Simmons, T. T. Hien, B. Wills, J. Farrar, S. Gupta, Immunological serotype interactions and their effect on the epidemiological pattern of dengue, *Proc. R. Soc. B Biol. Sci.* **276**, 2541–2548 (2009).
68. C. E. Wagner, M. Hooshyar, R. E. Baker, W. Yang, N. Arinaminpathy, G. Vecchi, C. J. E. Metcalf, A. Porporato, B. T. Grenfell, Climatological, virological and sociological drivers of current and projected dengue fever outbreak dynamics in Sri Lanka, *J. R. Soc. Interface* **17** (2020), doi:10.1098/rsif.2020.0075.
69. R. Montesano-Castellanos, C. Ruiz-Matus, Vigilancia epidemiológica del dengue en México, *Salud Publica Mex.* **370**, S64–S76 (1995).
70. B. Briseño-García, H. Gómez-Dantés, E. Argott-Ramírez, R. Montesano, A. L. Vázquez-Martínez, S. Ibáñez-Bernal, G. Madrigal-Ayala, C. Ruíz-Matus, A. Flisser, R. Tapia-Conyer, Potential Risk for Dengue Hemorrhagic Fever: The Isolation of Serotype Dengue-3 in Mexico, *Emerg. Infect. Dis.* **2**, 133–135 (1996).

71. H. Gómez Dantés, J. A. Farfán-Ale, E. Sarti, Epidemiological Trends of Dengue Disease in Mexico (2000–2011): A Systematic Literature Search and Analysis, *PLoS Negl. Trop. Dis.* **8** (2014), doi:10.1371/journal.pntd.0003158.
72. F. Dzul-Manzanilla, F. Correa-Morales, A. Che-Mendoza, J. Palacio-Vargas, G. Sánchez-Tejeda, J. F. González-Roldan, H. López-Gatell, A. E. Flores-Suárez, H. Gómez-Dantes, G. E. Coelho, H. S. da Silva Bezerra, N. Pavia-Ruz, A. Lenhart, P. Manrique-Saide, G. M. Vazquez-Prokopec, Identifying urban hotspots of dengue, chikungunya, and Zika transmission in Mexico to support risk stratification efforts: a spatial analysis, *Lancet Planet. Heal.* **5**, e277–e285 (2021).
73. L. Fernandes-Matano, I. E. Monroy-Muñoz, H. D. Pardavé-Alejandre, L. A. Uribe-Noguez, M. de los A. Hernández-Cueto, T. Rojas-Mendoza, C. E. Santacruz-Tinoco, C. Grajales-Muñiz, J. E. Muñoz-Medina, Impact of the introduction of chikungunya and zika viruses on the incidence of dengue in endemic zones of mexico, *PLoS Negl. Trop. Dis.* **15**, 1–16 (2021).
74. S. Zárate, F. Hernández-Perez, B. Taboada, N. E. Martínez, S. L. Alcaráz-Estrada, O. del Moral, M. Yocupicio-Monroy, Complete genome of DENV2 isolated from mosquitoes in Mexico, *Infect. Genet. Evol.* **71**, 98–107 (2019).
75. E. Carrillo-Valenzo, R. Danis-Lozano, J. X. Velasco-Hernández, G. Sánchez-Burgos, C. Alpuche, I. López, C. Rosales, C. Baronti, X. de Lamballerie, E. C. Holmes, J. Ramos-Castañeda, Evolution of dengue virus in Mexico is characterized by frequent lineage replacement, *Arch. Virol.* **155**, 1401–1412 (2010).
76. E. Hernández-García, M. de L. Muñoz, R. E. David, G. Pérez-Ramírez, J. Navarrete-Espinosa, Á. Díaz-Badillo, E. Domínguez-de-la-Cruz, M. Moreno-Galeana, C. A. Brito-Carreón, Epidemiological implications of the genetic diversification of dengue virus (DENV) serotypes and genotypes in Mexico, *Infect. Genet. Evol.* **84**, 104391 (2020).
77. N. Li, Y. Feng, B. Vrancken, Y. Chen, L. Dong, Q. Yang, M. U. G. Kraemer, O. G. Pybus, H. Zhang, O. J. Brady, H. Tian, Assessing the impact of COVID-19 border restrictions on dengue transmission in Yunnan Province, China: an observational epidemiological and phylogenetic analysis, *Lancet Reg. Heal. - West. Pacific* **14**, 100259 (2021).
78. I. de D. y R. E. "Dr. M. M. B. InDRE, Lineamientos para la vigilancia por laboratorio del sengue y otras arbovirosis (2021).
79. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. De Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples, *Nat. Protoc.* **12**, 1261–1266 (2017).
80. S. C. Hill, J. N. De Vasconcelos, B. Gutierrez Granja, J. Thézé, D. Jandondo, Z. Neto, A. Luísa, M. Cândido, C. Clemente, S. Pereira, T. De Oliveira, O. G. Pybus, N. R. Faria, Early Genomic Detection of Cosmopolitan Genotype of Dengue Virus Serotype 2, Angola, 2018, *Emerg. Infect. Dis.* **25**, 784–787 (2019).

81. V. Fonseca, P. J. K. Libin, K. Theys, N. R. Faria, M. R. T. Nunes, M. I. Restovic, M. Freire, M. Giovanetti, L. Cuypers, A. Nowé, A. Abecasis, K. Deforche, G. A. Santiago, I. C. de Siqueira, E. J. San, K. C. B. Machado, V. Azevedo, A. M. Bispo-De Filippis, R. V. da Cunha, O. G. Pybus, A. M. Vandamme, L. C. J. Alcantara, T. de Oliveira, A computational method for the identification of dengue, zika and chikungunya virus species and genotypes, *PLoS Negl. Trop. Dis.* **13**, 1–15 (2019).
82. S. V. Edgerton, P. Thongsripong, C. Wang, M. Montaya, A. Balmaseda, E. Harris, S. N. Bennett, Evolution and epidemiologic dynamics of dengue virus in Nicaragua during the emergence of chikungunya and Zika viruses, *Infect. Genet. Evol.* **92**, 104680 (2021).
83. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability, *Mol. Biol. Evol.* **30**, 772–780 (2013).
84. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, R. Lanfear, E. Teeling, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era, *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
85. S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0, *Syst. Biol.* **59**, 307–321 (2010).
86. A. Rambaut, T. T. Lam, L. M. Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen), *Virus Evol.* **2**, 1–7 (2016).
87. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, *Virus Evol.* **4**, 1–5 (2018).
88. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics, *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
89. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence, *PLoS Biol.* **4**, 699–710 (2006).
90. M. A. R. Ferreira, M. A. Suchard, Bayesian analysis of elapsed times in continuous-time Markov chains, *Can. J. Stat.* **36**, 355–368 (2008).
91. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior summarization in Bayesian phylogenetics using Tracer 1.7, *Syst. Biol.* **67**, 901–904 (2018).
92. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots, *PLoS Comput. Biol.* **5** (2009), doi:10.1371/journal.pcbi.1000520.
93. V. N. Minin, M. A. Suchard, Counting labeled transitions in continuous-time Markov models of evolution, *J. Math. Biol.* **56**, 391–412 (2008).
94. M. U. G. Kraemer, A. Sadilek, Q. Zhang, N. A. Marchal, G. Tuli, E. L. Cohn, Y. Hswen, T. A. Perkins, D. L. Smith, R. C. Reiner, J. S. Brownstein, Mapping global variation in human mobility, *Nat. Hum. Behav.* **4**, 800–810 (2020).
95. T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, H. Tatlow, A global panel database of

pandemic policies (Oxford COVID-19 Government Response Tracker), *Nat. Hum. Behav.* **5**, 529–538 (2021).
96. N. S. D. Sahadeo, O. M. Allicock, P. M. De Salazar, A. J. Auguste, S. Widen, B. Olowokure, C. Gutierrez, A. M. Valadere, K. Polson-Edwards, S. C. Weaver, C. V. F. Carrington, Understanding the evolution and spread of chikungunya virus in the Americas using complete genome sequences, *Virus Evol.* **3**, 1–10 (2017).
97. J. Thézé, T. Li, L. du Plessis, J. Bouquet, M. U. G. Kraemer, S. Somasekar, G. Yu, M. de Cesare, A. Balmaseda, G. Kuan, E. Harris, C. hsi Wu, M. A. Ansari, R. Bowden, N. R. Faria, S. Yagi, S. Messenger, T. Brooks, M. Stone, E. M. Bloch, M. Busch, J. E. Muñoz-Medina, C. R. González-Bonilla, S. Wolinsky, S. López, C. F. Arias, D. Bonsall, C. Y. Chiu, O. G. Pybus, Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico, *Cell Host Microbe* **23**, 855-864.e7 (2018).
98. T. F. Kautz, E. E. Díaz-González, J. H. Erasmus, I. R. Malo-García, R. M. Langsjoen, E. I. Patterson, D. I. Auguste, N. L. Forrester, R. M. Sanchez-Casas, M. Hernández-Ávila, C. M. Alpuche-Aranda, S. C. Weaver, Chikungunya Virus as Cause of Febrile Illness Outbreak, Chiapas, Mexico, 2014, *Emerg. Infect. Dis.* **21**, 2070–2073 (2015).
99. N. Sahadeo, H. Mohammed, O. M. Allicock, A. J. Auguste, S. G. Widen, K. Badal, K. Pulchan, J. E. Foster, S. C. Weaver, C. V. F. Carrington, Molecular Characterisation of Chikungunya Virus Infections in Trinidad and Comparison of Clinical and Laboratory Features with Dengue and Other Acute Febrile Cases, *PLoS Negl. Trop. Dis.* **9**, 1–18 (2015).
100. C. Wang, S. Saborio, L. Gresh, M. Eswarappa, D. Wu, A. Fire, P. Parameswaran, A. Balmaseda, E. Harris, Chikungunya virus sequences across the first epidemic in Nicaragua, 2014-2015, *Am. J. Trop. Med. Hyg.* **94**, 400–403 (2016).
101. J. P. Carrera, Y. Díaz, B. Denis, I. Barahona de Mosca, D. Rodriguez, I. Cedeño, D. Arauz, P. González, L. Cerezo, L. Moreno, L. García, L. E. Sáenz, M. A. Atencio, E. Rojas-Fermin, F. Vizcaino, N. Perez, B. Moreno, S. López-Vergès, A. Valderrama, B. Armién, Unusual pattern of chikungunya virus epidemic in the Americas, the Panamanian experience, *PLoS Negl. Trop. Dis.* **11**, 1–23 (2017).
102. F. Riosmena, D. S. Massey, Pathways to El Norte: Origins, destinations, and characteristics of Mexican migrants to the United States, *Int. Migr. Rev.* **46**, 3–36 (2012).
103. D. S. Massey, Patterns of U.S. Migration from Mexico, the Caribbean, and Central America, *Migr. Int.* **2**, 05–39 (2003).
104. L. Faret, M. E. A. Téllez, L. H. Rodríguez-Tapia, Migration Management and Changes in Mobility Patterns in the North and Central American Region, *J. Migr. Hum. Secur.* **9**, 63–79 (2021).
105. M. U. G. Kraemer, R. C. Reiner, O. J. Brady, J. P. Messina, M. Gilbert, D. M. Pigott, D. Yi, K. Johnson, L. Earl, L. B. Marczak, S. Shirude, N. Davis Weaver, D. Bisanzio, T. A. Perkins, S. Lai, X. Lu, P. Jones, G. E. Coelho, R. G. Carvalho, W. Van Bortel, C. Marsboom, G. Hendrickx, F. Schaffner, C. G. Moore, H. H. Nax, L. Bengtsson, E. Wetter, A. J. Tatem, J. S. Brownstein, D. L. Smith, L. Lambrechts, S. Cauchemez, C. Linard, N. R. Faria, O. G. Pybus, T. W. Scott, Q. Liu, H. Yu, G. R. W.

Wint, S. I. Hay, N. Golding, Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus, *Nat. Microbiol.* **4**, 854–863 (2019).
106. G. Muñoz, X. Chourio, A. Rivière-Cinnamond, M. A. Diuk-Wasser, P. A. Kache, E. A. Mordecai, L. Harrington, M. C. Thomson, AeDES: a next-generation monitoring and forecasting system for environmental suitability of Aedes-borne disease transmission, *Sci. Rep.* **10**, 1–13 (2020).
107. P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, M. A. Suchard, Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2, *PLoS Pathog.* **10** (2014), doi:10.1371/journal.ppat.1003932.
108. N. R. Faria, A. C. Da Costa, J. Lourenço, P. Loureiro, M. E. Lopes, R. Ribeiro, C. S. Alencar, M. U. G. Kraemer, C. J. Villabona-Arenas, C. H. Wu, J. Thézé, K. Khan, S. E. Brent, C. Romano, E. Delwart, B. Custer, M. P. Busch, O. G. Pybus, E. C. Sabino, C. D. A. Neto, A. Mendrone-Junior, A. Bárbara Carneiro-Proietti, D. D. A. Sampaio, C. Lobo, L. Capuani, J. E. Ferreira, M. Oikawa, P. Losco Takecian, C. D. L. Oliveira, S. Kelly, T. T. Gonçalez, D. Brambilla, C. McClure, S. A. Glynn, Genomic and epidemiological characterisation of a dengue virus outbreak among blood donors in Brazil, *Sci. Rep.* **7**, 1–12 (2017).
109. I. Y. Amaya-Larios, R. A. Martínez-Vega, S. V. Mayer, M. Galeana-Hernández, A. Comas-García, K. J. Sepúlveda-Salinas, J. A. Falcón-Lezama, N. Vasilakis, J. Ramos-Castañeda, Seroprevalence of neutralizing antibodies against dengue virus in two localities in the state of Morelos, Mexico, *Am. J. Trop. Med. Hyg.* **91**, 1057–1065 (2014).
110. T. J. Hladish, C. A. B. Pearson, D. L. Chao, D. P. Rojas, G. L. Recchia, H. Gómez-Dantés, M. E. Halloran, J. R. C. Pulliam, I. M. Longini, Projected Impact of Dengue Vaccination in Yucatán, Mexico, *PLoS Negl. Trop. Dis.* **10**, 1–19 (2016).
111. C. P. Farrington, E. Miller, Vaccine trials, *Appl. Biochem. Biotechnol. - Part B Mol. Biotechnol.* **17**, 43–58 (2001).
112. N. De Maio, C. H. Wu, K. M. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, *PLoS Genet.* **11**, 1–22 (2015).
113. R. Beard, D. Magee, M. A. Suchard, P. Lemey, M. Scotch, Generalized linear models for identifying predictors of the evolutionary diffusion of viruses., AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci. 2014, 23–8 (2014).

6 Phylogenetic analysis of the 2018 epidemic of Chikungunya virus infection in Kassala, Sudan

Highlight: *High domestic sampling* over a limited time period restricts phylodynamic inference

This chapter includes a phylogenetic analysis of a set of sequences from a Chikungunya outbreak in Kassala, Sudan. The short sampling time during which the sequenced samples were collected, despite their considerable number, highlights the limitations of genomic epidemiology when the temporal sampling aspect is reduced.

This chapter presents an investigation of a 2018 CHIKV outbreak in Kassala, Sudan, carried out by the UK-PHRST and the Ministry of Health of Sudan. I conceived, performed, interpreted and wrote all the phylogenetic analyses in the published paper. This includes Figure 3, Supplementary Figure 1 and the *Genetic analysis* section in the following publication:

- Bower H, el Karsany M, Hussein Hussein AAA, Idriss MI, al Zain MA, Alfakiyousif MEA, Mohamed R., Mahmoud I, Albadri O, Mahmoud SAA, Abdalla OI, Eldigail M, Elagib N, Arnold U, **Gutierrez B** *et al*. 2021. *Kankasha in Kassala: a prospective observational cohort study of the clinical characteristics, epidemiology, genetic origin, and chronic impact of the 2018*

epidemic of Chikungunya virus infection in Kassala, Sudan. ***PLoS Neglected Tropical Diseases*** 15(4): e0009387.

PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

"Kankasha" in Kassala: A prospective observational cohort study of the clinical characteristics, epidemiology, genetic origin, and chronic impact of the 2018 epidemic of Chikungunya virus infection in Kassala, Sudan

Hilary Bower[1‡]*, Mubarak el Karsany[2,3‡], Abd Alhadi Adam Hussein Adam[4], Mubarak Ibrahim Idriss[5], Ma'aaza Abasher Alzain[6], Mohamed Elamin Ahmed Alfakiyousif[2], Rehab Mohamed[2], Iman Mahmoud[2], Omer Albadri[7], Suha Abdulaziz Alnour Mahmoud[8], Orwa Ibrahim Abdalla[8], Mawahib Eldigail[2], Nuha Elagib[2], Ulrike Arnold[1], Bernardo Gutierrez[9], Oliver G. Pybus[9], Daniel P. Carter[10], Steven T. Pullan[10], Shevin T. Jacob[11], Tajeldin Mohammedein Abdallah[4,8‡], Benedict Gannon[1‡], Tom E. Fletcher[11‡]

1 UK Public Health Rapid Support Team, London School of Hygiene & Tropical Medicine/Public Health England, London, United Kingdom, **2** National Public Health Laboratory, Federal Ministry of Health, Khartoum, Sudan, **3** Karary University, Omdurman, Sudan, **4** University of Kassala, Kassala, Sudan, **5** Laboratory Division, Kassala State Ministry of Health, Kassala, Sudan, **6** Communicable Disease Surveillance & Events Unit, Federal Ministry of Health, Khartoum, Sudan, **7** Health Emergency and Epidemic Control Directorate, Federal Ministry of Health, Khartoum, Sudan, **8** Kassala Teaching Hospital, Kassala, Sudan, **9** Department of Zoology, University of Oxford, Oxford, United Kingdom, **10** National Infection Service, Public Health England, Porton, United Kingdom, **11** Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

‡ HB and MK share first authorship on this work. TMA, BG and TEF are joint senior authors on this work.
* hilary.bower@lshtm.ac.uk

Citation: Bower H, el Karsany M, Adam AAAH, Idriss MI, Alzain MA, Alfakiyousif MEA, et al. (2021) "Kankasha" in Kassala: A prospective observational cohort study of the clinical characteristics, epidemiology, genetic origin, and chronic impact of the 2018 epidemic of Chikungunya virus infection in Kassala, Sudan. PLoS Negl Trop Dis 15(4): e0009387. https://doi.org/10.1371/journal.pntd.0009387

Editor: Brett M. Forshey, DoD - AFHSB, UNITED STATES

Received: September 23, 2020

Accepted: April 14, 2021

Published: April 30, 2021

Data Availability Statement: The authors confirm that all data underlying the findings are available on request from the LSHTM Data Compass Repository DOI: https://doi.org/10.17037/DATA.00002018. Study consent requires data be made available for ethically-approved research only: requests can be made here: https://datacompass.lshtm.ac.uk/cgi/request_doc?docid=10142 or by emailing researchdatamanagement@lshtm.ac.uk.

Abstract

Background

The public health impact of Chikungunya virus (CHIKV) is often underestimated. Usually considered a mild condition of short duration, recent outbreaks have reported greater incidence of severe illness, fatality, and longer-term disability. In 2018/19, Eastern Sudan experienced the largest epidemic of CHIKV in Africa to date, affecting an estimated 487,600 people. Known locally as Kankasha, this study examines clinical characteristics, risk factors, and phylogenetics of the epidemic in Kassala City.

Methodology/Principal findings

A prospective cohort of 102 adults and 40 children presenting with chikungunya-like illness were enrolled at Kassala Teaching Hospital in October 2018. Clinical information, socio-demographic data, and sera samples were analysed to confirm diagnosis, characterise illness, and identify viral strain. CHIKV infection was confirmed by real-time reverse transcription-PCR in 84.5% (120/142) of participants. Nine (7.5%) CHIKV-positive participants had concurrent Dengue virus (DENV) infection; 34/118 participants (28.8%) had a positive

All study tools and documentation can be downloaded directly without request from https://doi.org/10.17037/DATA.00002018Gene sequences are available on Genbank (accession numbers: MW161364-161460.

Funding: HB was funded by UK Aid from the Department of Health and Social Care (https://www.gov.uk/government/collections/official-development-assistance-oda–2, Grant No. IS-RRT-1015-001) via the UK Public Health Rapid Support Team Research Programme (Grant No. EPIDZK 3819). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Rapid Diagnostic Test for *Plasmodium falciparum;* six (5.0%) had haemorrhagic symptoms including two children with life-threatening bleeding. One CHIKV-positive participant died with acute renal injury. Age was not associated with severity of illness although CHIKV-infected participants were younger (p = 0.003).

Two to four months post-illness, 63% of adults available for follow-up (30) were still experiencing arthralgia in one or more joints, and 11% remained moderately disabled on Rapid3 assessment. Phylogenetic analysis showed all CHIKV sequences from this study belonged to a single clade within the Indian Ocean Lineage (IOL) of the East/Central/South African (ECSA) genotype. History of contact with an infected person was the only factor associated with infection (p = 0.01), and likely related to being in the same vector environment.

Conclusions/Significance

Vulnerability to CHIKV remains in Kassala and elsewhere in Sudan due to widespread *Aedes aegypti* presence and mosquito-fostering household water storage methods. This study highlights the importance of increasing awareness of the severity and impact of CHIKV outbreaks, and the need for urgent actions to reduce transmission risk in households.

Author summary

Chikungunya is an arboviral disease transmitted to humans by infected mosquitoes and characterised by fever and arthralgia. Although it is generally considered a short self-limiting infection, long term sequelae and severe disease are increasingly recognised. In 2018/19, Eastern Sudan experienced the largest epidemic of Chikungunya in Africa to date, affecting approximately 500,000 people. We undertook a prospective hospital-based cohort study of patients presenting with undifferentiated febrile illness in Kassala city, Sudan, supported by next-generation sequencing. We confirmed that CHIKV was the dominant pathogen, with positive CHIKV RT-PCR in 85% of patients presenting during the 7-day study period. Dengue virus was also circulating with nine CHIKV RT-PCR-positive patients co-infected, and we identified high rates of *Plasmodium falciparum* infection and CHIKV/*P.falciparum* co-infection. Genetic sequencing confirmed Indian Ocean Lineage of the East/Central/South African CHIKV genotype. A quarter of participants available for follow-up (16/60, 26.6%) reported being admitted to hospital including two children with haemorrhage, reflecting the severe phenotype linked to this genotype. Increased understanding of the health and economic burden of Chikungunya is needed, and recognition that severe and occasionally fatal infection exists. With widespread presence of *Ae. aegypti* and household water storage practices that encourage mosquito breeding, timely actions will be essential to prevent further large outbreaks.

Introduction

Chikungunya is often considered a mild illness and its public health impact underestimated in Africa and the Middle East despite the fact that greater severity and long-term sequelae have been reported increasingly in recent years.[1–3]

First isolated in Tanzania in 1952,[4] Chikungunya virus (CHIKV) was implicated in rural outbreaks and sporadic cases across Africa until 1980 when it effectively disappeared. In 2004, however, it reappeared with a vengeance in a pandemic that spread from Kenya to the Indian Ocean islands and Asia, causing millions of infections.[5,6] In central Africa, sizeable CHIKV outbreaks in cities signalled a major shift in transmission to urban settings.[2,3,7–9]

Early symptoms of CHIKV are similar to many tropical febrile illnesses but it is differentiated by debilitating arthralgia, often bilaterally and in multiple joints. Other symptoms include headache, gastrointestinal problems, fatigue, asthenia, peripheral oedema and conjunctivitis.[10] Although most cases improve after 1–2 weeks, recent outbreaks have seen higher incidences of severe illness, including sepsis and cardiac, renal, neurological, skin and ocular manifestations, and of longer-term effects including persistent pain, rheumatic symptoms, depression, and mood and sleep disorders.[11–16] A meta-analysis of chronic symptoms studies in 15 outbreak countries found 43% of cases had not recovered at three months and 21% were unrecovered at 12 months. Long term neurodevelopmental delays have been reported in infants symptomatically infected through vertical transmission.[17,18] Treatment for Chikungunya involves supportive care and anti-inflammatory drugs, but as yet there is no antiviral treatment nor vaccine.[19]

Recent outbreaks have also challenged the view that CHIKV fatality is rare. Case mortality of 10–48% was reported in the Reunion Island outbreak among cohorts of hospitalised patients.[13,16] Central nervous system infections with fatal outcomes have been reported in the Americas.[20] Other locations have documented substantial rises in all-cause deaths during outbreak periods.[20–23]

Hypotheses to explain these changes include persistent immune activation triggered by viral debris, improved surveillance, and genetic mutation, particularly a new variant known as the Indian Ocean Lineage (IOL) which, since the 2004–2006 pandemic, has been linked to greater severity and the highest rates of non-recovery.[24] IOL has also been linked to increased risk of outbreak as some forms contain a mutation that allows CHIKV to use the temperate-dwelling *Aedes albopictus* mosquito as a vector as well the tropical *Aedes aegypti*.[5,24]

In low-resource settings, outbreaks of undifferentiated febrile illness (UFI) are both common and a diagnostic challenge. The Federal Republic of Sudan has seen 12 major outbreaks of UFI since 2012, frequently associated with haemorrhage and high case fatality ratios [25] and a range of high-consequence infectious diseases are endemic or epidemic in Sudan. Crimean-Congo haemorrhagic fever was recently identified as a common pathogen in a retrospective investigation of outbreaks in Darfur.[25] Yellow fever virus epidemics occur [26] and outbreaks of dengue fever are common.[27] The Federal Ministry of Health (FMoH) maintains a sentinel surveillance system and supports individual states with rapid response teams and enhanced diagnostic capability at the National Public Health Laboratory (NPHL) in Khartoum.

In late July 2018, the sentinel surveillance in Kassala State (Fig 1) detected an increased frequency of UFI, and CHIKV was identified by the NPHL in the blood sample of a traveller to Kassala city from nearby Red Sea State. In mid-September reports of cases with more severe symptoms, including haemorrhage, raised fears that another pathogen might be involved. Using a pre-prepared, locally and internationally-approved protocol, FMoH and Kassala Teaching Hospital (KTH), supported by Kassala University and the UK Public Health Rapid Support Team, deployed a small pre-trained study team of clinical, epidemiology and laboratory staff to investigate the outbreak syndrome, confirm the outbreak pathogen, and sequence the outbreak strain.

Fig 1. Study location: *Kassala state, located in Eastern Sudan, 600 km the capital city, Khartoum.* covers an area of 42,282 km^2 and has a population 1.8 million inhabitants. Kassala City is the State capital with a population of ~ 400,000. Kassala Teaching Hospital is the State tertiary hospital and provides service for all patients referred from health centres and rural hospitals. The source of the map is: https://commons.wikimedia.org/w/index.php?curid=16093732.

https://doi.org/10.1371/journal.pntd.0009387.g001

Methods

Ethics statement

Ethical approval was granted by the FMoH Technical Review Board, and Ethics Committees of Karary University, Khartoum (03 August 2017) and the London School of Hygiene & Tropical Medicine (Ref: 11930). All participants or parents/guardians of children aged under 18 years provided written informed consent and assent was sought from children aged 12–18.

Participant recruitment

The study was a prospective hospital-based observational cohort of consecutive patients presenting to the Kassala Teaching Hospital medical and paediatrics outpatient departments between 10th and 16th October 2018. Recruitment criteria were individuals of any age or sex with symptoms consistent with the KTH case definition for Chikungunya, namely history of fever or a measured fever of >37.5˚C (tympanic), and at least three of the following clinical features: headache, anorexia, lethargy, aching muscles or joints, breathing difficulties, vomiting, diarrhoea, stomach pain, difficulty swallowing, and hiccups, or bleeding of any kind. Patients with a positive malaria RDT were enrolled if they fit the study syndrome in case of co-infection. Participants were enrolled at presentation and blood samples taken immediately. Epidemiological information, clinical symptoms at presentation and subsequent laboratory results were recorded on standardised pro-formas. The follow-up component took place 70–120 days after enrolment (Jan-Feb 2019) depending on the availability of participants. Adults who responded to phone follow-up provided convalescent samples and completed the WHO-validated Routine Assessment of Patient Index Data 3 (RAPID3) disability and pain survey.

[28] Children were not asked to return due to common reluctance to allow blood-draw from healthy children, but outcome and duration of hospital stay were confirmed with parents by phone. All contactable participants were given their RT-PCR test results.

The study protocol (available at https://doi.org/10.17037/DATA.00002018) was based on the International Severe Acute Respiratory and Emerging Infections Consortium (ISARIC) Clinical Characterization Protocol for Severe Emerging Infections and adapted to context.[29] The sample size of 140 cases was informed by literature[3] and feasible recruitment in the urgent time frame (7 days).

Laboratory analysis

Haematology (Mindaray 3000 Plus, China), biochemistry analysis (Biosystem BTS 310, Germany), and malaria rapid diagnostic tests (*P. falciparum* RDT, SD Bioline, USA) were performed on all participants. Plasma samples were frozen at -80˚C and transported under liquid nitrogen to the NPHL in Khartoum where nucleic acid was extracted (QIAamp Viral RNA Mini kit, Qiagen, Germany) and RT-PCR performed to detect CHIKV RNA (RealStar Chikungunya RT-PCR Kit 2.0, Altona Diagnostics, Germany) and DENV RNA (RealStar Dengue RT-PCR Kit 2.0, Altona Diagnostics, Germany) using Rotor Gene Q or Corbett RG6000 thermocyclers. Exposure to CHIKV and DENV was assayed by anti-chikungunya and anti-dengue IgM and IgG indirect ELISA (Euroimmun, Germany).

Due to limited availability of anti-chikungunya IgM ELISA kits, CHIKV RT-PCR-negative samples were prioritised for IgM testing, followed by 69 CHIKV RT-PCR-positives in order of recruitment. All samples were tested for CHIKV IgG. An Ebolavirus RT-PCR (RealStar Ebolavirus RT-PCR Kit 1.0, Altona Diagnostics, Germany) was performed on seven samples from patients with haemorrhagic symptoms, collected prior to the study to rule out this pathogen: all were negative.

Data management

All patient data were entered into encrypted software and analysed with STATA (StataCorp LLC, USA, V14.2). Descriptive statistics are expressed as medians and Inter-Quartile Range (IQR) or means and standard deviations (SD) for continuous variables, and frequencies and proportions for categorical variables. Fisher's exact, Chi^2, T-tests and Spearman's Rank Order coefficients were used to assess association and correlation related to CHIKV RT-PCR result. Significance was set at $p<0.05$. Only RT-PCR results were used in comparative analyses.

Genetic sequencing

Complementary DNA (cDNA) was prepared and Sequence Independent Single Primer Amplification (SISPA) performed prior to library preparation using Oxford Nanopore kits (SQK-LSK108). Sequencing on flow cells (FLO-MIN106) was done using the MinION (Oxford Nanopore Technologies) as previously described.[30] Thirty samples were sequenced at NPHL and analysed using Albacore 1.2 to basecall, Mash Screen [31] was used to identify the most closely matched genome on Genbank, and a reference guided genome assembly constructed as described previously.[30] The first seven sequences were used to construct a basic phylogenetic analysis using MEGA version 6 in order to identify the Kassala strain in Sudan. With FMoH permission, sample aliquots were also transferred to PHE Porton and sequenced using the same methods. The resulting viral genetic sequences were sent to Oxford University Department of Zoology for phylogenetic analysis.

Whole genome sequences were aligned with publicly available complete CHIKV genomes belonging to the ECSA genotype using MAFFT *v7.450* implemented in Geneious R8.[32] A

maximum likelihood (ML) phylogenetic tree was estimated using RAxML 8.0 [33] under a GTR substitution model with a Gamma model of among-site rate heterogeneity. The estimated ML phylogeny was midpoint rooted, and node support was assessed through 100 ML bootstrap replicates. The geographic location and amino acid identity at site E1-226 of each sequence were annotated in the tree to explore the phylogenetic distribution of those traits.

Results

Of 155 patients approached over seven days, 102 adults and 40 children were included in the analysis. Eleven patients or their guardians refused to participate, and two cases whose samples were lost were excluded. (Fig 2).

Participants were aged 4 months to 70 years (mean 27 years, SD 17.6); 47.9% were female, none were pregnant (Table 1). Almost half (46.1%, 65/141) were people likely to spend more time in a household compound (e.g. housewives, unemployed, retired, children under 5). Median household size was 8. Two-thirds of participants lived in brick/concrete houses with their own well and sanitation, the remainder lived in less permanent material structures with shared water and sanitation. A third kept animals in their compound.

Laboratory results

Of 142 participants, 120 were confirmed positive for CHIKV infection by RT-PCR, giving the case definition used a positive predictive value of 84.5%. CHIKV RT-PCR cycle threshold (Ct) values ranged from 12–37 (mean 27.3, SD 7.6); over a quarter (27.5%) had a Ct value corresponding with a high viral load (Ct<20) (Table 2). A significant positive correlation existed between Ct value and both the number of days between onset and recruitment/sampling (Spearman r = 0.37, p<0.001) and lymphocyte count (r = 0.23, p = 0.012). No other significant correlations were found. Prolonged viremia was observed in seven CHIKV RT-PCR-positives

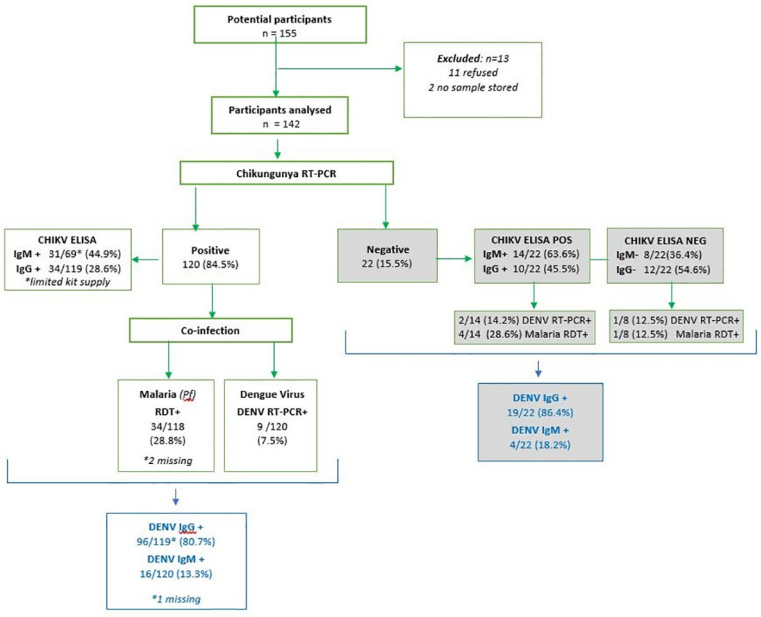


Fig 2. Flow diagram of study participants and virological findings: *CHIKV*: *Chikungunya; DENV*: *Dengue Virus*. ELISA IgM testing was limited by kit availability: CHIKV RT-PCR-negative samples were prioritised, followed 69 CHIKV RT-PCR positive samples in order of recruitment.

https://doi.org/10.1371/journal.pntd.0009387.g002

Table 1. Participant characteristics by Chikungunya (CHIKV) RT-PCR result.

	total	CHIKV positive (n = 120)	%	CHIKV negative (n = 22)	%
Sex					
Female	47.9%	56	*82.3*	12	*17.7*
Male	52.1%	64	*86.5*	10	*13.5*
Age (y), median [IQR]		22	*14–40*	37	*21–54*
Age Group (y) (n = 141)					
<2	7	7	*5.9*	0	*0.0*
2–4	6	5	*4.2*	1	*4.5*
5–14	24	22	*18.5*	2	*9.1*
15–29	51	46	*38.7*	5	*22.7*
30–49	33	26	*21.8*	7	*31.8*
50+	20	13	*10.9*	7	*31.8*
Occupation (n = 141)					
< 5 child	13	12	*10.0*	1	*4.8*
Schoolchild/student	49	43	*35.8*	6	*28.6*
Housewife	31	26	*21.7*	5	*23.8*
Farmer/Outdoor worker	15	12	*10.0*	3	*14.3*
Professional/Business	7	7	*5.8*	0	*0.0*
Health worker	5	3	*2.5*	2	*9.5*
Retired/unemployed	21	17	*14.2*	4	*19.0*
Median household size (IQR) n = 102	8	8	*6–11*	7	*5–9*

Notes: Missing data: age 1, occupation 1, # in household 40. IQR: Interquartile Range

Most participants who provided residence location (69%, 80/116) came from Kassala City sectors 2, 3, 4 and 5 on the banks of the seasonal River Gash, coinciding with areas of greatest flooding in the 2018 rainy season, and with highest case reports during the epidemic. A further 26% (30/116) were from rural areas up to 1.5 hours' drive away.

https://doi.org/10.1371/journal.pntd.0009387.t001

who presented 10 days after disease onset. Of 91 participants tested for CHIKV IgM antibodies, 45 were positive including 14 participants who were negative on CHIKV RT-PCR (Fig 2). There are several possible explanations for the IgM positive but RT-PCR negative cases: an acute or recent infection outside the viraemic phase of infection and therefore outside the diagnostic window for RT-PCR, prior CHIKV infection but unrelated to this outbreak (as IgM may persist for months [34]) or a false assay result.

Nine CHIKV RT-PCR-positive participants (7.5%) were co-infected with DENV, defined as being positive on DENV RT-PCR. A further 14 CHIKV RT-PCR-positive participants had evidence of DENV IgM antibodies, potentially indicating current or recent DENV infection, or due to cross reactivity of other flavivirus infections. Of all samples (CHIKV-positive and negative), 81.6% (115/141) were DENV IgG-positive suggesting previous infection with DENV or cross-reactive flaviviruses. Of the 118 CHIKV RT-PCR-positive participants with a *falciparum* malaria RDT result, 28.8% were positive, with the highest proportion in those aged 15–29 years. No participants were co-infected with all three pathogens.

Among the 69 CHIKV RT-PCR-positive participants tested for CHIKV IgM, 31 (44.9%) were positive. Of the 119 CHIKV RT-PCR tested for CHIKV IgG, 34 (28.6%) were positive (Fig 2). Of the 22 CHIKV RT-PCR-negative participants, 14 (63.6%) were CHIKV-IgM positive and 10 (45.5%) were CHIKV IgG positive. Participants who were CHIKV-positive on both RT-PCR and IgM ELISA presented later in illness course than those who were CHIKV RT-PCR-positive and IgM-negative ($p<0.001$). This finding was replicated with CHIKV

Table 2. Clinical symptoms and cycle threshold (Ct) values of participants at presentation by CHIKV RT-PCR result (n = 142).

	CHIKV positive		CHIKV negative	
Days from onset to presentation (median)	2	*IQR 1–4*	2	*IQR 1–5*
Symptoms at presentation		*%*		*%*
Fever/history of fever	117	*97.5*	22	*100.0*
Any bleeding	6	*5.0*	1	*4.8*
Headache	97	*80.8*	17	*85.0*
Joint pain	93	*77.5*	15	*75.0*
Fatigue	85	*70.8*	17	*81.0*
Muscle pain	66	*55.0*	16	*76.2*
Back pain	55	*45.8*	9	*42.9*
Loss of appetite	46	*38.3*	6	*28.6*
Vomiting	45	*37.5*	8	*38.1*
Abdominal pain	32	*26.7*	8	*38.1*
Lack of strength	25	*20.8*	4	*100.0*
Dizziness	18	*15.0*	3	*14.3*
Chest pain	17	*14.2*	3	*15.0*
Cough	17	*14.2*	2	*9.5*
Rash	13	*10.8*	2	*9.5*
Diarrhoea	11	*9.2*	2	*10.0*
Dysphagia	9	*7.5*	3	*15.0*
Dehydration	9	*7.5*	1	*33.3*
Shortness of breath	7	*5.8*	2	*10.0*
Conjunctivitis	7	*5.8*	3	*15.0*
Sore throat	6	*5.0*	0	*0.0*
Confusion	3	*2.5*	0	*0.0*
RT-PCR cycle threshold (Ct)				
High (Ct <20)	33	*27.5*	-	
Moderate (Ct 20–29)	24	*20.0*	-	
Low (Ct 30–38)	63	*52.5*	-	
Other pathology				
Malaria RDT positive	34/118	*28.8*	5/22	*22.7*
DENV positive (RT-PCR)	9/120	*7.5*	3/22	*13.6*

Notes: RDT Rapid Diagnostic Test. Missing data: fever 3; onset date 3; malaria RDT 2.

https://doi.org/10.1371/journal.pntd.0009387.t002

RT-PCR & IgG-positive patients who presented later than CHIKV IgG-negative patients (p = 0.002).

Clinical presentation

Median delay from symptom onset to presentation was 2 days (IQR 1–4, n = 139) with no difference by CHIKV diagnosis or age (Table 2).

Most common symptoms at presentation among CHIKV RT-PCR-positive participants were current or history of fever (97.5%, 2.5% data unrecorded), headache (88.2%), fatigue (82.5%), muscle, joint and back pain (66.0%, 83.8%, 50.9% respectively), loss of appetite (40.4%) and vomiting (40.2%). Patients who were found to be CHIKV RT-PCR-negative had similar presenting symptoms. Five percent (6/120) of CHIKV RT-PCR-positive participants reported bleeding, including haematemesis (4), oral bleeding (2), epistaxis (3), petechiae (1), haemoptysis (1) and melaena (1). None of those with haemorrhagic symptoms were co-

infected with DENV but one–a 9-year-old child with haematemesis–was co-infected with malaria. Eighteen of 78 CHIKV RT-PCR-positive adults (23.1%) were hypotensive (systolic blood pressure <100 mmHg) and seven (9.0%) were tachycardic (pulse >100), one of whom was co-infected with DENV. Only one person was both tachycardic and hypotensive.

Information on duration of admission was collected at follow-up. Of the 30 child guardians and 30 adult participants who responded to follow-up, 16 (26.6%) reported being admitted. All were CHIKV RT-PCR positive and none were coinfected with DENV, but 60% (7/12 children, 2/4 adults) had a positive *falciparum* malaria RDT. Children were more likely to be admitted than adults (40.0% vs 13.3%, Fisher's p = 0.04) and for longer: mean 4.7 days (sd 4.1) compared to 2.3 days (sd 1.3) for adults. Both adults and children were more likely to be admitted if they were malaria positive (p = 0.002), and children if they were bleeding (p = 0.054). One of the four children admitted for seven or more days was malaria positive.

In analysis of blood chemistry, a significant difference was found in mean haematocrit levels between CHIKV RT-PCR-positive (mean 35.9, SD = 6.27) and negative participants (mean 38.9, SD = 7.41, p = 0.05). There were no other significant differences in blood chemistry. Heart rate and blood pressure were similar in both groups. In the CHIKV RT-PCR-positive cohort median haematology and biochemistry results were all within normal ranges (Table 3). Leucopenia (white blood count < 4 x 10^9/L) was observed in 29% (35/120) and lymphopenia (< 1 x 10^9/L) observed in 50.7% (61/120). Elevated AST (>40IU/L) was observed in 22.5% (27/120) with six participants at levels> 100 IU/L (max 259 IU/L). ALT levels >51 IU/L were seen in seven participants with four recording levels >100 (max 150 IU/l). Acute kidney injury was observed in one fatal case. Platelet counts < 100 x 10^9/L were recorded in 23/120 CHIKV RT-PCR-positive participants, of whom 14/23 (60.9%) were also *falciparum* malaria RDT-positive and none were DENV PCR-positive. Platelet counts < 50 x 10^9/L were observed in six participants, three of whom also had positive malaria RDT.

Compared to those with a single infection, CHIKV/DENV co-infected participants had higher mean systolic blood pressure (p = 0.03) haemoglobin (p = 0.01), haematocrit (p = 0.03),

Table 3. Mean clinical parameters of participants at presentation by CHIKV RT-PCR result.

	CHIKV positive	SD	*n*	CHIKV negative	SD	*n*
Heart rate (beats/min)	**91.2**	19.2	*114*	**83.5**	15.8	*21*
Respiratory rate	**20.4**	2.9	*12*	**18.0**	n/a	*1*
Systolic pressure	**102.9**	20.2	*78*	**106.3**	12.5	*20*
Diastolic pressure	**68.0**	13.1	*78*	**70.2**	7.5	*20*
Haemoglobin (g/dL)	**11.6**	2.2	*120*	**12.5**	2.5	*22*
Haematocrit (%)*	**35.9**	6.3	*119*	**38.9**	7.4	*22*
WBC (10^3/uL)	**6.1**	3.1	*120*	**4.8**	2.2	*22*
Lymphocytes (10^3/uL)	**2.1**	2.4	*118*	**2.9**	6.6	*22*
Platelets (10^9/L)	**227.0**	143.3	*120*	**190.6**	131.0	*22*
Nitrogen oxide	**3.8**	2.0	*113*	**2.9**	1.7	*22*
Urea (mg/dL)	**24.8**	14.9	*119*	**26.1**	11.3	*22*
Creatinine (mg/dL)	**0.8**	0.7	*119*	**0.9**	0.2	*22*
Bilirubin	**0.7**	1.5	*119*	**0.9**	1.3	*22*
AST (IU/L)	**36.0**	33.9	*119*	**28.2**	17.4	*22*
ALT (IU/L)	**23.1**	23.4	*118*	**21.5**	14.4	*22*
Albumin	**14.9**	6.8	*69*	**17.6**	6.4	*17*

Notes: WBC White Blood Cells; AST Aspartate aminotransferase; ALT Alanine aminotransferase.

*Significant difference p = 0.049

https://doi.org/10.1371/journal.pntd.0009387.t003

bilirubin (p = 0.002) and albumin (p = <0.001). They were also more likely to present with back pain than those with CHIKV alone (p = 0.003).

Severe or fatal disease

The study recruited one case with a fatal outcome–a male teacher aged 52 transferred to KTH 13 days after onset of symptoms. He was CHIKV RT-PCR-positive with a Ct of 30 and no coinfection. At admission he had petechiae and minor bruising but no other evidence of bleeding. He died from sepsis complicated by disseminated intravascular coagulation and acute kidney injury (Blood/Urea/Nitrogen (BUN) 154 mg/dL, creatinine 7.8 mg/dL, AST 259 IU/L, ALT 59 IU/L, platelets 111 x 10^9/L).

Two children *in extremis* with significant haemorrhage were recruited at the paediatric department. Both were CHIKV RT-PCR-positive without coinfection. The 16-year-old was admitted on day 2 of illness with a history of fatigue, headache, back pain, anorexia, oral mucosal haemorrhage, and profuse epistaxis. He was in shock (pulse 120, BP 84/50, respiratory rate 25) with a platelet count of 10 x 10^9/L, Hb 12.5 g/dL, WBC 10.5 x 10^3/uL, HCT 38.9%, AST 47 IU/L, ALT 24 IU/L, BUN 26 mg/dL, creatinine 0.8 mg/dL and CHIKV RT-PCR Ct 29.63. He was resuscitated with blood and platelet transfusion, underwent nasal packing, and was discharged after 15 days. The second child aged 7 years presented on day 0 of illness with epistaxis, haematemesis, melaena, bruising and anorexia. Her pulse was 130 and Hb 10.2 g/dL, WBC 4.6 x 10^3/uL, platelet count 20 x 10^9/L, HCT 30.7%, urea 22 mg/dL, creatinine 0.5 mg/dL, AST 175 IU/L, ALT 129 IU/L and CHIKV RT-PCR Ct of 26.8. She was also resuscitated with blood and platelets and discharged well after nine days' admission.

Factors associated with CHIKV positivity

The mean age of CHIKV RT-PCR-positive patients was significantly less than CHIKV RT-PCR-negative patients (p = 0.003). CHIKV RT-PCR-positives were also more likely to report having been in contact with someone who was ill (p = 0.01) but as CHIKV is not transmitted person-to-person, this finding likely relates to being in the same vector environment. No associations were found with any other exposure including mass gatherings, funerals/corpses, or contact with livestock, birds or rodents.

Convalescent follow-up

Thirty (29.7%) of the 102 adult participants were available for follow-up, 29 (28.7%) refused, and 42 (41.6%) could not be reached. Median time to follow-up for adults was 75 days (IQR 71–94) after enrolment; 52% were male. Parents of 30 of the 40 child participants responded to contact. All 60 participants followed-up were CHIKV RT-PCR-positive at baseline. Of adults followed-up 43.3% (13/30) had spent 1–2 weeks off work and 20.0% (6/30) 2 or more weeks (Table 4). Twenty-three of the 30 adult convalescent samples collected (76.7%) were IgG positive: 10 were IgG-positive in both acute and convalescent samples, 13 had seroconverted since recruitment, and seven remained IgG negative. Children were not sampled.

Chronic disability

Only adult follow-up participants were surveyed for ongoing disability; median age was 27 years (IQR 24–45). Of these, 63.3% (19/30) were still experiencing arthralgia, most frequently in the knee (46.7% 14/30), ankle (36.7% 11/30) wrist (30.0% 9/30) and shoulder (26.7% 8/30). Asked to rank pain and general wellbeing over the previous week on a scale where 10 was the most pain/most unwell, respondents reported a median pain score of 4 (IQR 2–5) and median

Table 4. Impact of CHIKV infection on work and disability 70–120 days after acute infection.

	Adults (n = 101)	*%/IQR*	Children (n = 40)	*%/IQR*	p (Fisher)
Follow-up response rate	30	*29.7%*	30	*75.0%*	
Median days to follow-up	75	*IQR 71–94*	-	-	
Hospital care (n = 30)					
Discharged the same day	26	*86.7%*	18	*60.0%*	p = 0.039
Admitted for at least 1 day	4	*13.3%*	12	*40.0%*	
Median days admitted	2	*IQR 1.5–3*	3	*IQR 2–7*	
Absence from work					
< 1 week	11	*36.7%*			
1–2 weeks	13	*43.3%*			
3–4 weeks	4	*13.3%*			
>4 weeks	2	*6.7%*			
RAPID3 disability score (n = 28)					
Near remission	8	*28.6%*			
Low disability	17	*60.7%*			
Moderate disability	3	*10.7%*			
Current presence of pain (n = 30)	19	*63.3%*			
Location of pain *(multiple possible)*					
Knee	14	*46.7%*			
Ankle	11	*36.7%*			
Wrist	9	*30.0%*			
Shoulder	8	*26.7%*			
Fingers	7	*23.3%*			
Median pain perception score	4	*IQR 2–5*			
Median wellbeing score	7	*IQR 3–8*			

Notes: Pain perception and wellbeing are scored on a 0–10 scale with 10 being the worst pain and feeling the worst. IQR: Inter Quartile Range

https://doi.org/10.1371/journal.pntd.0009387.t004

wellbeing score of 7 (IQR 3–8) (Table 4). A third (36.6%, 11/30) indicated their pain was greater or the same as during the acute phase of their illness, and even those with less pain than before reported continuing to take painkillers, anti-inflammatories, corticosteroids, or all three, for their condition. In total, 60% continued to take medication.

Of the 28 respondents who completed the RAPID3 disability survey, three-quarters reported a continuing effect on day-to-day activities, including difficulties getting out of bed (28%), walking 2 miles (35%), and getting in and out of a vehicle (17%), while 10–15% reported difficulty with activities such as turning on taps, bending and sleeping well. Rapid3 scaling indicated that, 70–120 days after their acute illness, 10.7% (3/28) of respondents were experiencing moderately severe disability and 60.7% (17/28) low severity disability. The remaining eight (28.6%) respondents were considered nearly recovered (Table 4). Neither age, sex, level of viremia on admission, or pre-existing conditions such as diabetes, hypertension and osteoarthritis, were associated with level of residual disability, or duration of absence from work.

Genetic analysis

Virus genome sequencing found that all CHIKV RT-PCR-positive samples belonged to a single monophyletic cluster in the Indian Ocean Lineage (IOL) of the ECSA genotype of CHIKV. (Fig 3) Between 84–95% (mean 93%) genome coverage was achieved at least x20 depth across

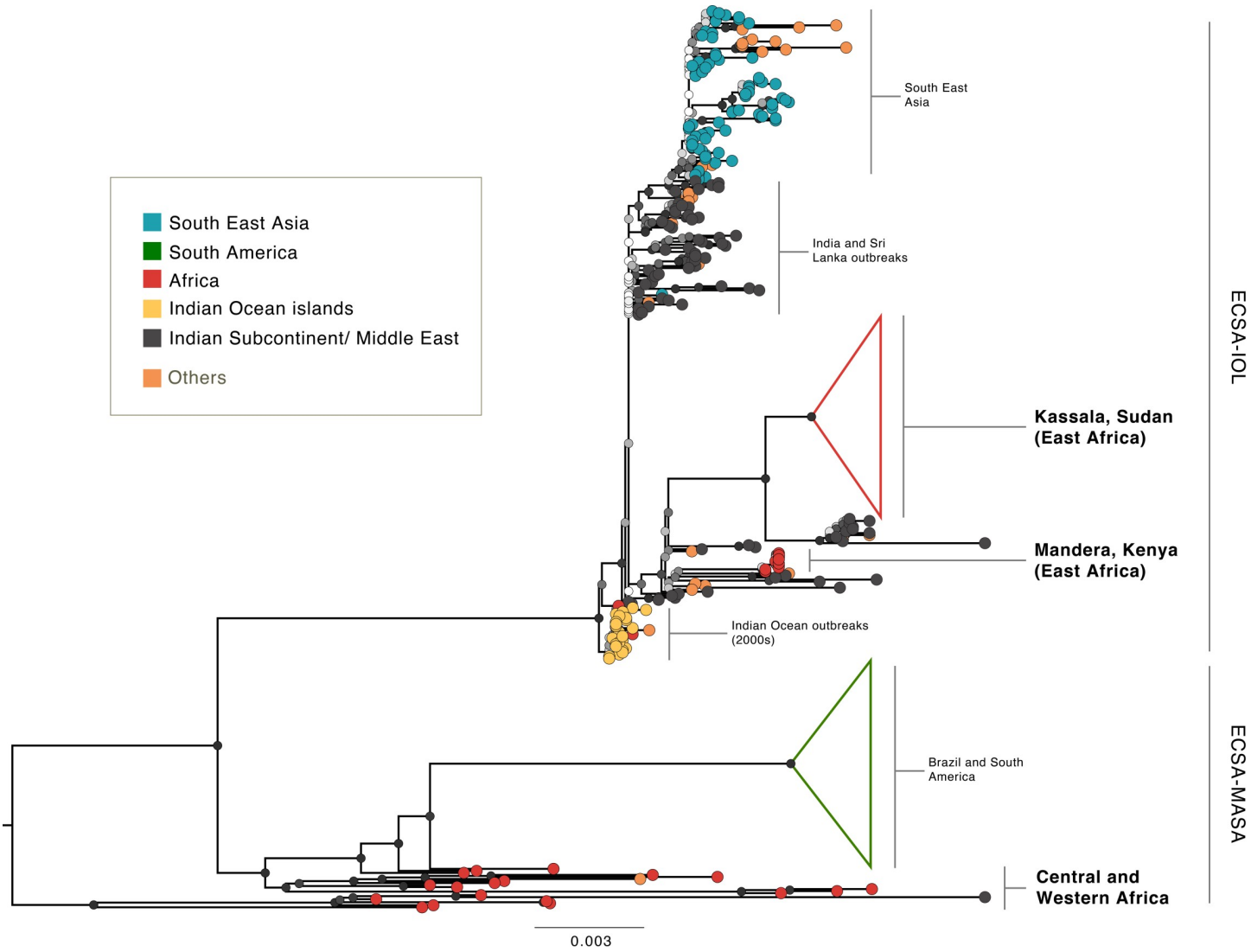


Fig 3. Maximum likelihood phylogenetic tree of the ECSA genotype and its two distinct lineages showing the location of the Kassala 2018/19 strain: ***Maximum likelihood phylogenetic tree of the East/Central/South African (ECSA) genotype and its two distinct lineages: the Indian Ocean Lineage (IOL) and the Middle African/ South American (MASA) lineage.*** The monophyletic Kassala epidemic clade groups with sequences that originate from the Middle East and India and represents a distinct lineage from other African CHIKV variants. Node support values were evaluated using 100 bootstrap replicates and are illustrated using colour at nodes (with white representing 0 and black representing 100). The 2016 Mandera outbreak also represents a possible introduction from the Indian subcontinent and appears to be an independent event from the Kassala epidemic. Neither the Mandera or Kassala outbreak contained the A226V albopictus adaptive variant.

https://doi.org/10.1371/journal.pntd.0009387.g003

all samples (Genbank Accession Nos. MW161364-161460). This clade is distinct from those previously identified in Western and Central Africa, suggesting that the Kassala outbreak was caused by an independent introduction of an IOL strain into the region. This might have occurred via the Middle East or Indian Subcontinent since the Kassala sequences are closely related (i) to the Henan and Shivane variants identified in China and Hong Kong, which originated in returning travellers from India/Pakistan, and (ii) to a clade that includes sequences from a 2016 outbreak in Pakistan and other sequences from India and Bangladesh. Absence of the A226V mutation—a mutation associated with the capacity of the virus to infect *Ae. Albopictus* [35,36]—confirms *Ae. aegypti* as the most likely outbreak vector species.(S1 Text)

Discussion

Our phylogenetic analysis suggests the unprecedented Chikungunya epidemic that took place in the Eastern states of Sudan in 2018–19 was caused by an independent introduction of a CHIKV IOL virus, of the same lineage responsible for the outbreaks observed worldwide during 2004–2007 and linked to more severe manifestations of CHIKV disease.[24]

Over the nine months of the epidemic, Kassala State and Red Sea State reported 48,763 cases of chikungunya. Seroprevalence surveys in other outbreak locations, however, suggest only 2–15% of cases are captured by surveillance and reporting systems.[37–40] A conservative estimate of 10% of cases captured suggests some 487,600 cases occurred in the Kassala and Red Sea State epidemic, making it the largest outbreak of Chikungunya in Africa and the Middle East to date. There are multiple reasons to suspect substantial under-reporting and under-diagnosis in this outbreak including: use of a sentinel surveillance system normally intended to raise an alert rather than capture every case (one study reported a ratio of 67 symptomatic cases for every one picked up by the sentinel system [40]), population reluctance to seek medical attention at hospitals due to cost and limited treatment options, limited diagnostic resources, absence of data collection in private clinics, and the range of CHIKV illness severity.

Made possible by the established research collaboration between the FMoH and the UK Public Health Rapid Support Team on outbreak response, this study utilised pre-approved protocols, pre-trained staff and, for the first time within Sudan, next generation sequencing technology in-country. We confirmed that CHIKV was the dominant pathogen responsible for the outbreak, that DENV was also circulating with 1 in 5 CHIKV positive patients co-infected and, importantly, that there was a high frequency of *falciparum* malaria transmission and of CHIKV/malaria co-infection. The 14 cases with chikungunya RT-PCR negative and ELISA IgM positive results were not unexpected given that the diagnostic window for IgM and RT-PCR are different (IgM antibodies emerge later during the infection and last for longer [34] than the viraemic phase when RNA is identified by RT-PCR). Other possible explanations are prior CHIKV infection but unrelated to this outbreak (as IgM may persist for months [34]) or a false assay result. As IgM/IgG results cannot confirm a diagnosis they were not used to define Chikungunya infection in any analyses.

The cohort recruited were young (median age 27) and largely presented with non-specific febrile symptoms plus poly-articular joint pain. Nevertheless, based on our follow-up cohort, a substantial proportion—one in four—were admitted to hospital, reflecting the impact of the high frequency of malaria co-infection (60%) as well as the severe phenotype observed in a subset of patients, including the potential for haemodynamic disturbance in adults. As we have described, 5% of patients with CHIKV (*P. falciparum* and DENV negative) had evidence of bleeding and loss of haemostasis. Two children in particular were gravely unwell with life-threatening haemorrhage associated with severe thrombocytopenia and need for blood product resuscitation, while one adult died of multi-organ failure including acute kidney injury.

Severe and fatal CHIKV infection has been reported previously during outbreaks challenging the misconception that it is only a mild self-limiting diseases.[13,41] Increased risk of severe disease has been suggested in extremes of age (neonates and $>$ 65 years) and those with underlying medical problems including immunosuppression.[10,15] Severe disease has been associated with neurological complications such as acute flaccid paralysis and Guillain-Barré syndrome.[42–44] Brazil, Reunion Island and India all reported higher case fatality and excess deaths during their epidemics, [12,13,16,21,22] though fatality is still considered rare, occurring in less than 1 in 1000 cases.

Although we did not find associations between severe illness and age in our study, or with chronic illness, due to our age-limited follow-up, the findings described by previous studies are consistent with the severe phenotype of CHIKV we observed in Kassala, including a one

confirmed CHIKV RT-PCR-positive death and life-threatening bleeding due to loss of haemostasis with thrombocytopenia. The high frequency of *P*. *falciparum* infection detected underpins the importance of both malaria diagnostics and public messaging in malaria-endemic settings when chikungunya outbreaks occur.

Our study was biased towards more severe cases due to hospital-based recruitment and had a low response rate in follow-up where we also lacked a control group. There was a low reported sensitivity of 76.7% (23/30) for the CHIKV IgG ELISA in the convalescent samples tested but it was not possible to determine whether this was due to assay failures or other factors, in part, due to the inability to retest original acute samples to verify results.

However, our findings that 63% of respondents had persistent pain three to four months after acute illness are similar to those of a case-controlled study on La Reunion Island. This study found 53% of IOL-strain CHIKV-positive participants reported twice as much pain compared to controls 12 months after their illness.[45] Our findings were also similar to those of a meta-analysis of chronic disability across all CHIKV strains which found on average 52% of IOL-strain patients were still disturbed by symptoms at three months, compared to 39% of Asian lineage and 14% of ECSA group patients.[24]

Our finding that the only significant risk factor for contracting CHIKV was proximity to a case suggests identifying and supporting vector control measures acceptable to the community should be a priority for public health authorities seeking to prevent future CHIKV and other *Aedes*-transmitted epidemics in Sudan. Entomological research carried out by Kassala University in 2014/15 showed high *Ae*. *aegypti* density across the city and a high proportion of households storing water in their compounds in un-protected clay pots.[46–48] In 2018 the situation was exacerbated by intense seasonal rainfall which displaced 50,000 residents and flooded many areas, a situation repeated in 2020.

Measures such as improved water infrastructure, covered water storage containers, and increased population awareness of the importance of preventing mosquitoes from breeding in household compounds are important.[46] Heightened vector surveillance is also needed to monitor for presence of *Ae*. *albopictus*, which could prompt further outbreaks of different CHIKV strains, and to investigate the potential role in the sylvatic cycle of the large population of Macaques monkeys which live close to residential areas.

Finally, but importantly, it is essential that State Ministries of Health throughout Sudan enhance their risk communication and public information strategies around Chikungunya to underline that severe and occasionally fatal infection exists. With widespread presence of *Ae*. *aegypti* and a similar water storage practices throughout Sudan, timely official and community actions will be needed to prevent another large outbreak in the near future.

ENDS

Supporting information

S1 STROBE checklist.
(DOC)

S1 Text. Phylogenetic tree showing viral adaptation to different vector species.
(DOCX)

Acknowledgments

Disclaimer

The study team appreciates the collaboration of the leadership of the University of Kassala, Kassala Teaching Hospital, Kassala State Ministry of Health, Federal Ministry of Health

Directorate of Health and Epidemics Control Directorate, and the National Public Health Laboratory. We also appreciate the efforts of the International Severe Acute Respiratory and Emerging Infection Consortium to create generic protocols including the Clinical Characterisation Protocol for Severe Emerging Infections from which our study protocol was adapted. [29] The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health and Social Care.

Author Contributions

Conceptualization: Hilary Bower, Mubarak el Karsany, Oliver G. Pybus, Steven T. Pullan, Shevin T. Jacob, Benedict Gannon, Tom E. Fletcher.

Data curation: Hilary Bower, Abd Alhadi Adam Hussein Adam, Iman Mahmoud, Benedict Gannon.

Formal analysis: Hilary Bower, Bernardo Gutierrez, Oliver G. Pybus, Tom E. Fletcher.

Funding acquisition: Hilary Bower.

Investigation: Mubarak el Karsany, Abd Alhadi Adam Hussein Adam, Mubarak Ibrahim Idriss, Ma'aaza Abasher Alzain, Mohamed Elamin Ahmed Alfakiyousif, Rehab Mohamed, Omer Albadri, Suha Abdulaziz Alnour Mahmoud, Orwa Ibrahim Abdalla, Mawahib Eldigail, Ulrike Arnold, Bernardo Gutierrez, Oliver G. Pybus, Daniel P. Carter, Steven T. Pullan, Tajeldin Mohammedein Abdallah, Benedict Gannon, Tom E. Fletcher.

Methodology: Hilary Bower, Mubarak el Karsany, Benedict Gannon, Tom E. Fletcher.

Project administration: Hilary Bower, Mubarak el Karsany, Tajeldin Mohammedein Abdallah, Tom E. Fletcher.

Resources: Hilary Bower, Mubarak el Karsany, Abd Alhadi Adam Hussein Adam, Nuha Elagib, Bernardo Gutierrez, Oliver G. Pybus, Benedict Gannon, Tom E. Fletcher.

Software: Bernardo Gutierrez, Daniel P. Carter.

Supervision: Hilary Bower, Mubarak el Karsany, Tajeldin Mohammedein Abdallah, Benedict Gannon, Tom E. Fletcher.

Visualization: Hilary Bower.

Writing – original draft: Hilary Bower, Tom E. Fletcher.

Writing – review & editing: Hilary Bower, Mubarak el Karsany, Abd Alhadi Adam Hussein Adam, Mubarak Ibrahim Idriss, Ma'aaza Abasher Alzain, Mohamed Elamin Ahmed Alfakiyousif, Rehab Mohamed, Iman Mahmoud, Omer Albadri, Suha Abdulaziz Alnour Mahmoud, Orwa Ibrahim Abdalla, Mawahib Eldigail, Nuha Elagib, Ulrike Arnold, Bernardo Gutierrez, Oliver G. Pybus, Daniel P. Carter, Steven T. Pullan, Shevin T. Jacob, Tajeldin Mohammedein Abdallah, Benedict Gannon, Tom E. Fletcher.

References

1. Bonney JH, Osei-Kwasi M, Adiku TK, Barnor JS, Amesiya R, Kubio C, et al. Hospital-based surveillance for viral hemorrhagic fevers and hepatitides in Ghana. PLoS neglected tropical diseases. 2013; 7(9): e2435. https://doi.org/10.1371/journal.pntd.0002435 PMID: 24069490
2. Russo G, Subissi L, Rezza G. Chikungunya fever in Africa: a systematic review. Pathogens and Global Health. 2020; 114(3):111–9. https://doi.org/10.1080/20477724.2020.1743087 PMID: 32191166

3. Humphrey JM, Cleton NB, Reusken CBEM, Glesby MJ, Koopmans MPG, Abu-Raddad LJ. Urban Chikungunya in the Middle East and North Africa: A systematic review. PLoS neglected tropical diseases. 2017; 11(6):e0005707. https://doi.org/10.1371/journal.pntd.0005707 PMID: 28651007
4. Robinson MC. An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1955; 49(1):28–32. https://doi.org/10.1016/0035-9203(55)90080-8 PMID: 14373834
5. Ng LC, Hapuarachchi HC. Tracing the path of Chikungunya virus—evolution and adaptation. Infect Genet Evol. 2010; 10(7):876–85. https://doi.org/10.1016/j.meegid.2010.07.012 PMID: 20654736
6. Rossi G, Karki S, Smith RL, Brown WM, Ruiz MOH. 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. https://doi.org/10.1016/j.sste.2018.06.002 PMID: 30390927
7. Leroy EM, Nkoghe D, Ollomo B, Nze-Nkogue C, Becquart P, Grard G, et al. Concurrent chikungunya and dengue virus infections during simultaneous outbreaks, Gabon, 2007. Emerging infectious diseases. 2009; 15(4):591–3. https://doi.org/10.3201/eid1504.080664 PMID: 19331740
8. Staples JE, Breiman RF, Powers AM. Chikungunya fever: an epidemiological review of a re-emerging infectious disease. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2009; 49(6):942–8. https://doi.org/10.1086/605496 PMID: 19663604
9. Pezzi L, Diallo M, Rosa-Freitas MG, Vega-Rua A, Ng LFP, Boyer S, et al. GloPID-R report on chikungunya, o'nyong-nyong and Mayaro virus, part 5: Entomological aspects. Antiviral research. 2020; 174:104670. https://doi.org/10.1016/j.antiviral.2019.104670 PMID: 31812638
10. van Aalst M, Nelen CM, Goorhuis A, Stijnis C, Grobusch MP. Long-term sequelae of chikungunya virus disease: A systematic review. Travel medicine and infectious disease. 2017; 15:8–22. https://doi.org/10.1016/j.tmaid.2017.01.004 PMID: 28163198
11. Paixão ES, Teixeira MG, Rodrigues LC. Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases. BMJ Global Health. 2018; 3(Suppl 1):e000530. https://doi.org/10.1136/bmjgh-2017-000530 PMID: 29435366
12. Burt FJ, Chen W, Miner JJ, Lenschow DJ, Merits A, Schnettler E, et al. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. The Lancet Infectious Diseases. 2017; 17(4):e107–e17. https://doi.org/10.1016/S1473-3099(16)30385-1 PMID: 28159534
13. Rajapakse S, Rodrigo C, Rajapakse A. Atypical manifestations of chikungunya infection. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2010; 104(2):89–96. https://doi.org/10.1016/j.trstmh.2009.07.031 PMID: 19716149
14. Mehta R, Gerardin P, de Brito CAA, Soares CN, Ferreira MLB, Solomon T. The neurological complications of chikungunya virus: A systematic review. Reviews in medical virology. 2018; 28(3):e1978–e. https://doi.org/10.1002/rmv.1978 PMID: 29671914
15. Economopoulou A, Dominguez M, Helynck B, Sissoko D, Wichmann O, Quenel P, et al. Atypical Chikungunya virus infections: clinical manifestations, mortality and risk factors for severe disease during the 2005–2006 outbreak on Réunion. Epidemiology and Infection. 2009; 137(4):534–41. https://doi.org/10.1017/S0950268808001167 PMID: 18694529
16. Lemant J, Boisson V, Winer A, Thibault L, Andre H, Tixier F, et al. Serious acute chikungunya virus infection requiring intensive care during the Reunion Island outbreak in 2005–2006. Crit Care Med. 2008; 36(9):2536–41. https://doi.org/10.1097/CCM.0b013e318183f2d2 PMID: 18679124
17. Contopoulos-Ioannidis D, Newman-Lindsay S, Chow C, LaBeaud AD. Mother-to-child transmission of Chikungunya virus: A systematic review and meta-analysis. PLoS neglected tropical diseases. 2018; 12(6):e0006510. https://doi.org/10.1371/journal.pntd.0006510 PMID: 29897898
18. Gérardin P, Sampériz S, Ramful D, Boumahni B, Bintner M, Alessandri J-L, et al. Neurocognitive Outcome of Children Exposed to Perinatal Mother-to-Child Chikungunya Virus Infection: The CHIMERE Cohort Study on Reunion Island. PLoS neglected tropical diseases. 2014; 8(7):e2996. https://doi.org/10.1371/journal.pntd.0002996 PMID: 25033077
19. Innovations CfEP. The world needs a chikungunya vaccine [Available from: https://cepi.net/news_cepi/the-world-needs-a-chikungunya-vaccine/.
20. Lima STSd, Souza WMd, Cavalcante JW, da Silva Candido D, Fumagalli MJ, Carrera J-P, et al. Fatal outcome of chikungunya virus infection in Brazil. Clinical Infectious Diseases. 2020. https://doi.org/10.1093/cid/ciaa1038 PMID: 32766829
21. Mavalankar D, Shastri P, Bandyopadhyay T, Parmar J, Ramani KV. Increased mortality rate associated with chikungunya epidemic, Ahmedabad, India. Emerg Infect Dis. 2008; 14(3):412–5. https://doi.org/10.3201/eid1403.070720 PMID: 18325255

22. Josseran L, Paquet C, Zehgnoun A, Caillere N, Le Tertre A, Solet J-L, et al. Chikungunya disease outbreak, Reunion Island. Emerging infectious diseases. 2006; 12(12):1994–5. https://doi.org/10.3201/eid1212.060710 PMID: 17354339

23. Freitas ARR, Donalisio MR, Alarcón-Elbal PM. Excess Mortality and Causes Associated with Chikungunya, Puerto Rico, 2014–2015. Emerging infectious diseases. 2018; 24(12):2352–5. https://doi.org/10.3201/eid2412.170639 PMID: 30277456

24. Paixao ES, Rodrigues LC, Costa M, Itaparica M, Barreto F, Gerardin P, et al. Chikungunya chronic disease: a systematic review and meta-analysis. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2018; 112(7):301–16. https://doi.org/10.1093/trstmh/try063 PMID: 30007303

25. Bower H, El Karsany M, Alzain M, Gannon B, Mohamed R, Mahmoud I, et al. Detection of Crimean-Congo Haemorrhagic Fever cases in a severe undifferentiated febrile illness outbreak in the Federal Republic of Sudan: A retrospective epidemiological and diagnostic cohort study. PLoS neglected tropical diseases. 2019; 13(7):e0007571. https://doi.org/10.1371/journal.pntd.0007571 PMID: 31291242

26. Markoff L. Yellow fever outbreak in Sudan. The New England journal of medicine. 2013; 368:689–91. https://doi.org/10.1056/NEJMp1300772 PMID: 23387798

27. World Health Organisation. Dengue Fever—Republic of the Sudan 2019 [Available from: https://www.who.int/csr/don/22-november-2019-dengue-sudan/en/.

28. American College of Rheumatology. Routine Assessment of Patient Index Data 3 (RAPID3) [Available from: https://www.rheumatology.org/Portals/0/Files/RAPID3%20Form.pdf.

29. International Severe Acute Respiratory and Emerging Infections Consortium, World Health Organisation. ISARIC-WHO Clinical Characterization Protocol for Severe Emerging Infections (v. 3 2014) [Available from: https://isaric.tghn.org/protocols/clinical-characterization-protocol/.

30. Kafetzopoulou LE, Efthymiadis K, Lewandowski K, Crook A, Carter D, Osborne J, et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2018; 23(50):1800228. https://doi.org/10.2807/1560-7917.ES.2018.23.50.1800228 PMID: 30563591

31. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. Genome biology. 2019; 20(1):232. https://doi.org/10.1186/s13059-019-1841-x PMID: 31690338

32. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

33. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

34. Chelluboina S, Robin S, Aswathyraj S, Arunkumar G. Persistence of antibody response in chikungunya. Virusdisease. 2019; 30(3):469–73. https://doi.org/10.1007/s13337-019-00534-5 PMID: 31803816

35. Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential. PLOS Pathogens. 2007; 3(12):e201. https://doi.org/10.1371/journal.ppat.0030201 PMID: 18069894

36. Lindh E, Argentini C, Remoli ME, Fortuna C, Faggioni G, Benedetti E, et al. The Italian 2017 Outbreak Chikungunya Virus Belongs to an Emerging Aedes albopictus-Adapted Virus Cluster Introduced From the Indian Subcontinent. Open forum infectious diseases. 2018; 6(1):ofy321–ofy.

37. Sergon K, Yahaya AA, Brown J, Bedja SA, Mlindasse M, Agata N, et al. Seroprevalence of Chikungunya virus infection on Grande Comore Island, union of the Comoros, 2005. Am J Trop Med Hyg. 2007; 76(6):1189–93. PMID: 17556634

38. Sergon K, Njuguna C, Kalani R, Ofula V, Onyango C, Konongoi LS, et al. Seroprevalence of Chikungunya virus (CHIKV) infection on Lamu Island, Kenya, October 2004. Am J Trop Med Hyg. 2008; 78(2):333–7. PMID: 18256441

39. Sissoko D, Malvy D, Giry C, Delmas G, Paquet C, Gabrie P, et al. Outbreak of Chikungunya fever in Mayotte, Comoros archipelago, 2005–2006. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2008; 102(8):780–6. https://doi.org/10.1016/j.trstmh.2008.02.018 PMID: 18400240

40. Renault P, Solet JL, Sissoko D, Balleydier E, Larrieu S, Filleul L, et al. A major epidemic of chikungunya virus infection on Reunion Island, France, 2005–2006. Am J Trop Med Hyg. 2007; 77(4):727–31. PMID: 17978079

41. Suhrbier A. Rheumatic manifestations of chikungunya: emerging concepts and interventions. Nature Reviews Rheumatology. 2019; 15(10):597–611. https://doi.org/10.1038/s41584-019-0276-9 PMID: 31481759

42. Sá PKdO, Nunes MM, Leite I, Campelo M, Leão C, Souza JdC, et al. Chikungunya virus infection with severe neurologic manifestations: report of four fatal cases. Revista da Sociedade Brasileira de Medicina Tropical. 2017; 50:265–8. https://doi.org/10.1590/0037-8682-0375-2016 PMID: 28562768

43. Wielanek AC, Monredon JD, Amrani ME, Roger JC, Serveaux JP. Guillain-barré syndrome complicating a chikungunya virus infection. Neurology. 2007; 69(22):2105. https://doi.org/10.1212/01.wnl.0000277267.07220.88 PMID: 18040016

44. Singh SS, Manimunda SP, Sugunan AP, Sahina, Vijayachari P. Four cases of acute flaccid paralysis associated with chikungunya virus infection. Epidemiology and infection. 2008; 136(9):1277–80. https://doi.org/10.1017/S0950268807009739 PMID: 18634716

45. Soumahoro M-K, Gérardin P, Boëlle P-Y, Perrau J, Fianu A, Pouchot J, et al. Impact of Chikungunya Virus Infection on Health Status and Quality of Life: A Retrospective Cohort Study. PLOS ONE. 2009; 4 (11):e7800. https://doi.org/10.1371/journal.pone.0007800 PMID: 19911058

46. Ahmed R, Hassan S, Abdallah K, Enan M. Breeding and Resting Behaviour of Aedes aegypti in Indoor and Outdoor Environment in Kassala City, Sudan 2014/2015. Health Science Journal. 2019; 13 (5: 672).

47. Ahmed R, Hassan S. Seasonal indices of Aedes aegypti (Diptera: Culicidae) in an urban area of Kassala City, Sudan, 2014–2015. Europ Acad Res. 2019; V1(10).

48. Ahmed R, Hassan S, Elrahman A. Climatic Factors Affecting Density of Aedes aegypti (Diptera: Culicidae) in Kassala City, Sudan 2014/2015. Asploro Journal of Biomedical and Clinical Case Reports. 2019; 2(2):58–68.

7 Discussion

Despite having developed relatively recently, the fields of phylodynamics and genomic epidemiology have seen tremendous progress over the last decade. The COVID-19 pandemic further catapulted the application and utility of these tools to higher levels of impact, as they are well adapted to study fast-evolving, emerging RNA viruses, such as SARS-CoV-2. This, combined with some the of largest virus genomic surveillance programmes established to date, has generated copious amounts of data from locations worldwide. Some recent technical advances have been driven by challenges particular to SARS-CoV-2, such as data generation and management, computational efficiency, and the flexibility to adapt to a dynamic pandemic in real-time (e.g., *(1–3)*).

The results in this thesis highlight the applicability of phylogenetic methods to reconstruct and evaluate the transmission and evolution of emerging viruses, including SARS-CoV-2 and other well studied pathogens. Chapter 2 describes the evolutionary genetic history of OROV and explores the effects of natural selection on the virus' genome as it spread across South America. The results show how the viral surface glycoprotein is affected by natural selection (both at the sequence level and through the emergence of novel N-linked glycosylation sites) and suggest that the virus engages in frequent reassortment, as inferred from sampled and unsampled genomic data (the latter

becoming evident through the discrepant root ages of the time-calibrated phylogenies of the individual genome segments). Chapter 3 focuses on the spatial transmission dynamics of SARS-CoV-2 in Ecuador and shows that the first year of the epidemic in Ecuador was driven by multiple introductions that rapidly spread within the country. The substantial number of introductions that did not appear to lead to further transmission (i.e., singleton sequences) observed in the two most populous provinces in the country (one of them featuring the highest excess mortality rate during 2020) suggest that these likely accounted for a substantial proportion of these introductions, especially during the early months of the pandemic. Large numbers of singletons correspond to either poorly sampled transmission lineages or true imports without onward transmission, emphasising the frequency at which viral importations took place on both the Guayas and Pichincha provinces.

Also concerning SARS-CoV-2, Chapter 4 investigates the emergence and circulation of a recombinant lineage in North America that spread across the USA, Mexico and Central America over several months. Given the limited amount of viral genetic diversity, these analyses also utilise information about deletions shared by the lineages under investigation to determine the likely parental lineages and the resulting recombinant lineage, now classified as the XB lineage in the Pango nomenclature *(4)*. Chapter 5 integrates genomic and epidemiological data to explore the transmission patterns of two important arboviruses in Mexico, CHIKV and DENV, and reconstructs their epidemic and spatial dissemination patterns. For this study, new genome sequences of CHIKV and DENV-2 were generated and analysed in context of epidemiological

surveillance and human mobility data. The results show that a southern region of Mexico frequently acts as a port of entry for different arboviruses into the country. Plausible sources for the viruses are identified, and I show that new arbovirus introductions are followed by changes in virus epidemiological trends over the following months or years. Finally, Chapter 6 shows that the 2018 outbreak of CHIKV in Kassala, Sudan, derived from a single introduction of the IOL genotype, which had previously circulated in the Indian Ocean territories and India before being re-introduced into Africa on at least two independent occasions (once in Kenya in addition to the Kassala introduction).

A common feature of these data sets and analyses is the presence of incomplete sampling. The vast heterogeneity in genome sampling across both temporal and spatial scales is one of the well-documented shortcomings of the current SARS-CoV-2 genomic surveillance efforts; while some coordinated national efforts have produced very large numbers of virus genomes for some countries, entire regions of the world have suffered from sparse genomic coverage *(5)*. This is exemplified by the relatively low number of genome sequences in the Ecuador SARS-CoV-2 data set (160 genomes), which I conclude fail to represent the early weeks of the epidemic in the country as the high weekly excess mortality data from the country during March and April does not correlate with the scarce number of genome sequences generated during these months (Chapter 3). It is also likely that the data set used to identify the XB linage suffers from some sampling gaps, explaining why genomes from the recombinant lineage were sampled earlier than genomes from both parental lineages. This is likely due to the relative frequency of the lineages compared to other circulating variants (Chapter 4). Beyond

SARS-CoV-2, the OROV data set comprises a maximum of 149 genomes (for the most frequently sequenced segment) from samples collected over 60 years of viral circulation in the region (Chapter 2). The genomic coverage of DENV-2 in Mexico (Chapter 5) and of CHIKV in both Mexico and Sudan (Chapter 5 and Chapter 6) provide a better representation of the viral circulation in their respective countries, but the phylogenetic analyses in both cases highlight the plausible sampling gaps in other countries that could be associated with the importation and circulation of the viruses.

7.1 From case identification to genomic sequencing

Sequencing viruses in an epidemic context relies on the identification of individual cases of the disease they cause. For this reason, the study of epidemics naturally brings forth practical challenges in data generation and collection. Outbreak investigations seldom allow for the design of a sampling scheme that guarantees that the data being produced will provide a complete and unbiased representation of the population at large. The notion that data imperfectly represents the phenomenon being studied, a key assumption of many methods of statistical inference, can be further amplified when practical issues of data collection in real-world scenarios arise. Furthermore, the unexpected nature of outbreaks means that the deployment of resources to perform data collection can be delayed, leaving irrecoverable information gaps.

It’s important to understand the process of data generation and how this ultimately affects phylodynamic inference, which begins with case identification. This is achieved through diagnostic testing of infected individuals, which is contingent on testing

capacity and distribution. For viral infections, this normally means laboratory capability, access to specialised equipment and a constant supply of reagents. Home tests as implemented during the COVID-19 pandemic (e.g., lateral flow tests, LFTs) are an exception to this observation, but some of the limitations of their deployment are similar to those of laboratory tests. The interplay between accessibility (i.e., are there sufficient healthcare providers to investigate all infection cases across space and time, and can all patients with suspected infection timely access to healthcare?) and capacity (i.e., are there enough technical personnel, home tests such as LFTs, reagents, laboratory equipment and other resources available to perform accurate diagnostic tests and report their results in a timely manner?) defines the baseline at which epidemiological data can be generated. Once diagnostics have been deployed, the attribution of metadata to individual records can add layers of actionable information to individually-diagnosed cases. Most importantly the demographic details of the person being diagnosed, the key dates of disease progression and the details of past movements can provide valuable information for efficient modelling of transmission dynamics (Annex 1; *(6)*). This has recently been characterised under a formal framework under the WHO Guidance for surveillance of SARS-CoV-2 variants *(7)*; at the highest priority level of expected metadata, the sampling strategy by which samples for sequencing are collected is identified as an important field. Establishing whether samples were part of random sampling or focused outbreak investigations can expand the potential epidemiological links between sequences and aid in the interpretation of downstream analyses.

This diagnostics pipeline is also the backbone of genomic data generation, which relies on the presence of sufficient viral genomic material in a sample for efficient sequencing *(8, 9)*. Data incompleteness in a genomic epidemiology context can be less serious than for traditional epidemiological data sets because phylodynamic analysis methods inherently draw information and establish links among samples from the genetic sequences themselves. As samples are not entirely independent observations but rather partially correlated due to shared ancestry, the links between samples are established by inferring the unobserved ancestors. However, sampling imbalances and gaps can still introduce bias into phylodynamic inference: the omission of important taxa (i.e., unsampled cases) can result in lineages not being identified, which is significant if the characteristics of such lineages contribute to the broader epidemic process. For example, when the goal of the analysis is to describe the spatiotemporal spread, missing cases that result in no forward transmission have less of an impact than missing cases that show where prevalent lineages were circulating. This issue is challenging in both systematics *(10)* and in genomic epidemiology, particularly given that it is unfeasible to predict the spatiotemporal distribution of individual lineages without any prior information *(11–13)*.

Related to this is the challenge of power calculations and estimating ideal sample sizes in genomic sampling, because the information content of a set of genomes is not necessarily defined directly by the number of sequences but rather by the genetic diversity represented in the data (i.e., the partial correlation between sequences arising from shared ancestry makes it inadequate to treat every new sequence in a data set as an

independent new observation). Even high rates of genomic sampling can still lead to insufficient information to accurately estimate key evolutionary parameters if the genomes contain little genetic diversity, a concept that has been proposed as the phylodynamic threshold *(14)*. The considerable genomic sampling in the 2018 CHIKV Kassala epidemic (Chapter 6) shows that large numbers of genome sequences, while increasing the certainty that a single introduction caused the entire epidemic, do not necessarily provide greater analytic power of the epidemic dynamics themselves.

7.2 Unsampled sequences in phylodynamics

A common feature of epidemic investigations is that sampling intensity fluctuates across space and time, as surveillance systems are implemented, scaled-up and extended to reach more remote areas *(15)*. This can lead to two possible outcomes: a poor correlation between number of sequences and number of new cases in an epidemic across time, and the unbalanced representation of specific geographic locations with respect to the spatial distribution of cases. Crucially, a combination of both situations can have strong effects on the resulting parameter estimates of the spatial diffusion process. Temporal sampling gaps can reduce the accuracy of the inference of when common ancestors occurred in the phylogeny, while missed sampling in specific geographic locations biases the estimate of the past locations where a pathogen circulated and from where it was exported to other destinations. Overall, inferred transitions between discrete locations, or the geographic accuracy and precision of inferred ancestors, can be considerably influenced by the sampling distribution *(11, 12)*.

This is evident from the analysis in Chapter 3, in which the long branches (on time-calibrated phylogenies) leading to some of SARS-CoV-2 transmission lineages in Ecuador serve as indicators of sampling gaps. In this analysis, some Ecuador transmission lineages exhibit long internal branches that connect older basal nodes to more recent terminal ones. These long branches possibly represent sampling gaps which, if filled with missing sequences from Ecuador, would help resolve the trajectory of these lineages within the country. The identity of the missing tips would provide additional information to identify spatial origins and routes of geographical spread, which is currently unfeasible using the available genomic data alone. In fact, the large number of infections (suggested by the high excess mortality rate) observed during the first weeks of the epidemic in Ecuador contrast with the low number of genomes from the same time period (Figure 1C, Chapter 3), indicating under-sampling during the early epidemic. This is highlighted by the data from the Guayas province, which greatly contributed to the excess mortality estimates for the country during the first wave. Despite the apparently good genomic representation of this province during 2020 (Figure 1B, Chapter 3), most of these samples were obtained after the severe first wave. Furthermore, a large number of singleton sequences can be observed in Guayas, suggesting plausible importations that led to limited to none forward transmission (Figure 3B, Chapter 3; Figure S2, Chapter 3). These two observations suggest that the genomic underrepresentation from the first wave in Guayas potentially misses the origin of a substantial portion of the viral genetic diversity circulating in the rest of the country.

Similar limitations are evident at longer timescales and for analyses featuring multiple discrete locations. The results in Chapter 5 show that DENV-2 was introduced into Mexico multiple times during the 2000s; interpretation of a DTA phylogegraphic model suggests that the source of these importations was Nicaragua, however it is likely that countries like Guatemala, Honduras and El Salvador also play a major role in seeding DENV-2 epidemics in Mexico. This hypothesis could be evaluated phylogenetically if better genomic coverage from these countries was available, but some evidence for it can be obtained indirectly. The identifiable role of the Southwestern region of Mexico as a location of entry for arboviruses (Figure 2 Chapter 5; Figure 4 Chapter 5) and the region's high connectivity with Guatemala, as measured through smartphone usage data (Figure 5 Chapter 5), suggests that Guatemala could act as a transit state through which lineages from Nicaragua pass and circulate on their way to being introduced into Mexico. A prediction from this hypothesis is that sequences from Guatemala would tend to be basal to the Mexico clusters, as is the case for the only two available Guatemalan DENV-2 sequences. Similarly, a more detailed analysis of the phylogenetic reconstruction of the spread of CHIKV around the Indian Ocean and its introduction into Sudan would test the possibility that the sequences from the 2018 Kassala epidemic derive from unsampled outbreaks in nearby regions such as the Middle East, rather than from the widely sampled epidemics in India (Figure 3 Chapter 6). More generally, the phenomenon uneven spatio-temporal sampling should be further explored by contrasting phylogenetic results with epidemiological and mobility data, and ideally by integrating these data sources.

7.3 Data contrast versus data integration

The examination of evidence from independent data sources can improve the confidence of the conclusions drawn in a range of research fields. This approach is a time-tested feature of scientific inquiry, and an important component of the use of phylogenetic methods to understand the spread and evolution of viruses, validate findings and open new research questions. Given that interactions among hosts play a significant role in the transmission of pathogens, epidemiological, behavioural and spatial data on the hosts are now recognised as important data sources that can contextualise genomic findings and to provide clearer views of pathogen transmission *(16)*. However, on a technical level, simple 'compare and contrast' approaches to independent data sources can limit the capacity to formally test hypotheses. An important limitation of Chapter 5, in which the phylogeographic patterns of DENV-2 and CHIKV spread are compared to human mobility patterns, is that the effect of mobility on viral spread could not be directly quantified since our smartphone use data (used to infer human mobility trends) does not overlap with the genomic sequence data sampling dates; at a more basic level, results can be interpreted in light of both sets of observations.

On the other hand, data integration can be a powerful way to effectively and jointly maximise the utility of both genomic and non-genomic data, such as with the incorporation of external information sources into phylogenetic spatial diffusion models *(11)*. A good example for this comes from the COVID-19 pandemic. SARS-CoV-2 genomic data was used to track the spread of the virus from its likely origin in Hubei,

China, into Europe and North America. However, the severe first waves in countries like Italy and Iran limited the generation of genomic data from these locations, which represented a gap in the reconstruction of the complete geographic spread of the virus during these early stages. In light of this, human mobility data was incorporated into phylodynamic models in the form of self-reported travel histories from sequenced patients; this solution yielded an improved resolution on the broad mobility patterns of specific lineages *(13, 17)*. In fact, multiple studies on the importation dynamics of the virus to different countries rely mostly on epidemiological and travel data rather to identify importation sources rather than genomic data as a way to circumvent inherent sampling biases and to confirm or inform phylogenetically inferred importations *(17–20)*. An alternative approach is the integration of other independent data sources for human mobility to inform the phylogeographic transition rates, either as explicit model priors *(21)* or through a GLM that directly informs the diffusion process *(13, 22–24)*.

7.4 Future work

Given that infectious disease epidemics can be spatially structured and conceived as complex networks or hierarchical metapopulations *(25)*, and that heterogeneous sampling across interconnected locations can bias epidemiological model parameter estimates *(11, 26)*, data generation practices across administrative divisions will affect the capacity to compare trends across countries and regions. The inference of epidemic dynamics also presents a major challenge to the generation of models that integrate diverse data generation regimes and accurately describe transmission dynamics over large areas *(27)*. The data-collection schemes and descriptions presented in the Annex

section (Annex 1) highlight the opportunities and challenges of data integration and showcase some of the data imbalances that can become problematic. For example, the estimation of the number of new pathogen importations derived from incidence data combined with mobility data *(18, 28)* can be influenced by inaccurate incidence measures from locations with poor epidemiological surveillance *(29)*. It is therefore important to consider the sources of information, their limitations in the local context and the comparability between estimates.

Among these limitations is the awareness of sampling intensity across time. Explicit models that account for changes in sampling intensity in phylogenetic inferential frameworks have been recently developed and evaluated *(30–32)*, and represent an exciting new avenue of research. Likewise, given the bias in the estimation of ancestral node locations in a phylogeny through spatial diffusion phylogeographic methods *(11, 12)*, new approaches that account for spatial sampling imbalances *(26)* can better accommodate for diverse genomic data sets and improve their inferential accuracy. These new models need to be further validated (particularly for their universal applicability and limitations) and, if successful, will be of particular importance for locations where sequencing capabilities can only scale-up to a limited extent. In this sense, the expansion of the current set of phylodyamic methods will be crucial to improve the usability of data with important sampling gaps, as these types of data sets will likely remain a prevalent source of information from settings where genomic data generation is challenging.

7.5 General conclusions

The results presented in this thesis present new insights into the transmission and evolution of different emerging viruses across different settings where sampling gaps are common, whether due to the opportunistic nature of sampling (as with the analysis of historical data sets from virus collections) or to the challenges of generating genome data in real time during an outbreak due to limited resources. Furthermore, it is evident that tracking the spread and evolution of viruses across countries adds additional challenges related to the differences in the capacity to generate genomic data from distinct locations. Nonetheless, the wealth of information provided by genomic data facilitates the study of important phenomena where sampling gaps themselves can be informative.

The genomic surveillance of emerging pathogens is turning into a public health priority for countries across the globe and is of particular importance for regions where the likelihood of new viral threats is high due to their geographical and climatic features, biodiversity and increasingly frequent contacts between humans and animals. Despite the remarkable advances in establishing sequencing capacity across the globe (accelerated by the COVID-19 pandemic) challenges to random, population level sampling (mostly tied to preparedness, institutional and resource limitations, as well as access to remote populations) mean that information gaps will likely remain a prevalent feature of genomic data sets. Alongside the optimisation of genomic surveillance itself (a task that should focus on currently underdeveloped regions), there needs to be further development and evaluation of methods that make better use of different information

sources and which utilise and synergise all kinds of epidemiological data, from traditional case data to pathogen genomes. For this purpose, the wealth of genomic data available for some regions can serve as an important validation tool: down-sampling large genomic data sets and evaluating the accuracy of inferential methods on these reduced data sets, compared to the full data set, can provide empirical means to test the effects of sampling on phylodynamic inference. This particular topic is an area of active research at the time of writing. In a widely interconnected world where epidemic threats become increasingly shared amongst all nations, this approach presents a way to develop broadly applicable tools for the individual needs of different countries and to improve our general knowledge of the emergence of new viruses.

7.6 References

1. S. Argimón, K. Abudahab, R. J. E. Goater, A. Fedosejev, J. Bhai, C. Glasner, E. J. Feil, M. T. G. Holden, C. A. Yeats, H. Grundmann, B. G. Spratt, D. M. Aanensen, Microreact: visualizing and sharing data for genomic epidemiology and phylogeography, *Microb. genomics* **2**, e000093 (2016).
2. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, NextStrain: Real-time tracking of pathogen evolution, *Bioinformatics* **34**, 4121–4123 (2018).
3. Á. O'Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J. T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. Du Plessis, D. Maloney, N. Medd, S. W. Attwood, D. M. Aanensen, E. C. Holmes, O. G. Pybus, A. Rambaut, Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool, *Virus Evol.* **7**, 1–9 (2021).
4. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, *Nat. Microbiol.* **5**, 1403–1407 (2020).
5. A. F. Brito, E. Semenova, G. Dudas, G. W. Hassler, C. C. Kalinich, M. U. G. Kraemer, S. C. Hill, Danish Covid-19 Genome Consortium, E. C. Sabino, O. G. Pybus, C. Dye, S. Bhatt, S. Flaxamn, M. A. Suchard, N. D. Grubaugh, G. Baele, N. R. Faria, Global disparities in SARS-CoV-2 genomic surveillance., *medRxiv Prepr. Serv. Heal. Sci.* , 1–24 (2021).

6. A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. G. Kraemer, J. Wu, K. Leung, G. M. Leung, Modelling COVID-19, *Nat. Rev. Phys.* **2**, 279–281 (2020).
7. WHO, *Guidance for surveillance of SARS-CoV-2 variants: interim guidance* (2021; https://apps.who.int/iris/rest/bitstreams/1361901/retrieve).
8. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. De Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples, *Nat. Protoc.* **12**, 1261–1266 (2017).
9. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar, *bioRxiv* , 1–19 (2018).
10. X. Hua, R. Lanfear, The influence of non-random species sampling on macroevolutionary and macroecological inference from phylogenies, *Methods Ecol. Evol.* **9**, 1353–1362 (2018).
11. A. Kalkauskas, U. Perron, Y. Sun, N. Goldman, G. Baele, S. Guindon, N. De Maio, Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk, *PLoS Comput. Biol.* **17**, 1–27 (2021).
12. N. De Maio, C. H. Wu, K. M. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, *PLoS Genet.* **11**, 1–22 (2015).
13. P. Lemey, S. L. Hong, V. Hill, G. Baele, C. Poletto, V. Colizza, Á. O'Toole, J. T. McCrone, K. G. Andersen, M. Worobey, M. I. Nelson, A. Rambaut, M. A. Suchard, Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2, *Nat. Commun.* **11**, 1–14 (2020).
14. S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, G. Baele, Temporal signal and the phylodynamic threshold of SARS-CoV-2, *Virus Evol.* **6**, 1–8 (2020).
15. A. F. Brito, E. Semenova, G. Dudas, G. W. Hassler, C. C. Kalinich, M. U. G. Kraemer, S. C. Hill, Danish Covid-19 Genome Consortium, E. C. Sabino, O. G. Pybus, C. Dye, S. Bhatt, S. Flaxamn, M. A. Suchard, N. D. Grubaugh, G. Baele, N. R. Faria, Global disparities in SARS-CoV-2 genomic surveillance., *medRxiv Prepr. Serv. Heal. Sci.* , 1–14 (2021).
16. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century, *Nat. Microbiol.* **4**, 10–19 (2019).
17. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America, *Science (80-.).* **370**, 564–570 (2020).

18. D. Da S. Candido, A. Watts, L. Abade, M. U. G. Kraemer, O. G. Pybus, J. Croda, W. De Oliveira, K. Khan, E. C. Sabino, N. R. Faria, Routes for COVID-19 importation in Brazil, *J. Travel Med.* **27**, 1–3 (2020).
19. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK, *Science (80-.).* **371**, 708–712 (2021).
20. M. Ghafari, B. Hejazi, A. Karshenas, S. Dascalu, A. Kadvidar, M. A. Khosravi, M. Abbasalipour, M. Heydari, S. Zeinali, L. Ferretti, A. Ledda, A. Katzourakis, Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran, *Epidemics* **36**, 100472 (2021).
21. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots, *PLoS Comput. Biol.* **5** (2009), doi:10.1371/journal.pcbi.1000520.
22. P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, M. A. Suchard, Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2, *PLoS Pathog.* **10** (2014), doi:10.1371/journal.ppat.1003932.
23. R. Beard, D. Magee, M. A. Suchard, P. Lemey, M. Scotch, Generalized linear models for identifying predictors of the evolutionary diffusion of viruses., *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.* **2014**, 23–8 (2014).
24. D. Magee, M. A. Suchard, M. Scotch, Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction, *PLoS Comput. Biol.* **13**, 1–19 (2017).
25. D. J. Watts, R. Muhamad, D. C. Medina, P. S. Dodds, Multiscale, resurgent epidemics in a hierarchical metapopulation model, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 11157–11162 (2005).
26. S. Guindon, N. De Maio, Accounting for spatial sampling patterns in Bayesian phylogeography, (2021), doi:10.1073/pnas.2105273118/-/DCSupplemental.Published.
27. M. U. G. Kraemer, S. V. Scarpino, V. Marivate, B. Gutierrez, B. Xu, G. Lee, J. B. Hawkins, C. Rivers, D. M. Pigott, R. Katz, J. S. Brownstein, Data curation during a pandemic and lessons learned from COVID-19, *Nat. Comput. Sci.* **1**, 9–10 (2021).
28. I. I. Bogoch, A. Watts, A. Thomas-Bachli, C. Huber, M. U. G. Kraemer, K. Khan, Potential for global spread of a novel coronavirus from China, *J. Travel Med.* **27**, 1–3 (2020).
29. A. Karlinsky, D. Kobak, Tracking excess mortality across countries during the covid-19 pandemic with the world mortality dataset, *Elife* **10**, 1–21 (2021).
30. K. V. Parag, L. Du Plessis, O. G. Pybus, Jointly inferring the dynamics of population size and sampling intensity from molecular sequences, *Mol. Biol. Evol.* **37**, 2414–2429 (2020).
31. M. D. Karcher, L. M. Carvalho, M. A. Suchard, G. Dudas, V. N. Minin, Estimating effective population size changes from preferentially sampled genetic sequences, *PLoS Comput. Biol.* **16**, 1–22 (2020).

32. L. A. Featherstone, F. Di Giallonardo, E. C. Holmes, T. G. Vaughan, S. Duchêne, Infectious disease phylodynamics with occurrence data, *Methods Ecol. Evol.* **12**, 1498–1507 (2021).

8 Appendices

Appendix 1 Related publications

I participated in a paper led by Emma Wise describing the generation of Oropouche virus data.

- Wise EL, Márquez S, Mellors J, Paz V, Atkinson B, **Gutierrez B** et al. 2020. Oropouche virus cases identified in Ecuador using an optimised qRT-PCR informed by metagenomic sequencing. **PLoS Neglected Tropical Diseases** 14(1): e0007897.

The methods used in Chapter 3 were inspired by a paper led by Louis du Plessis describing the importation and transmission lineage dynamics of SARS-CoV-2 in the UK. I contributed to the summarisation of transmission lineages in the UK, the compilation of travel restriction information in the UK and with sensitivity analyses for the importation model.

- du Plessis L*, McCrone JT*, Zarebski AE*, Hill V*, Ruis C*, **Gutierrez B**, Raghwani J *et al*. 2021. *Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK*. ***Science*** 371(6530): 708-712.

Additional research outputs from the SARS-CoV-2 genomic surveillance efforts in Ecuador were also published. I contributed to the following publications:

- Prado-Vivar B, Becerra-Wong M, Guadalupe JJ, Márquez S, **Gutierrez B**, Rojas-Silva P *et al*. 2021. *A case of SARS-CoV-2 reinfection in Ecuador*. ***The Lancet Infectious Diseases*** 21(6): e142.
- Márquez S, Prado-Vivar B, Guadalupe JJ, **Gutierrez B**, Becerra-Wong M, Jibaja M et al. 2020. Metagenome of a Bronchoalveolar Lavage Fluid Sample from a Confirmed COVID-19 Case in Quito, Ecuador, Obtained Using Oxford Nanopore MinION Technology. **Microbiology Resource Announcements** 9(41).

The DENV-2 sequencing protocols from Chapter 5 were designed and tested by SCH and reported in a paper to which I contributed with phylogenetic analyses.

- Hill SC, Vasconcelos JN, **Gutiérrez Granja B**, Thézé J, Jandondo D, Neto Z et al. 2019. Early Genomic Detection of Cosmopolitan Genotype of Dengue Virus Serotype 2, Angola, 2018. **Emerging Infectious Diseases** 25(4): 784.

Other papers relevant to this thesis
Work related to the collection and analysis of epidemiological data from the early COVID-19 pandemic.

- Meaver L, Amit M, Pepito VCF, **Gutierrez B**, Rawson T. 2021. Tracking changes in reporting of epidemiological data during the COVID-19 pandemic in Southeast Asia: an observational study during the first wave. **Frontiers in Public Health** 9: 2296-2565.

- Kraemer MUG, Scarpino SV, Marivate V, **Gutierrez B**, Xu B, Lee G et al. 2021. Data curation during a pandemic and lessons learned from COVID-19. **Nature Computational Science** 1: 9-10.
- Vespignani A, Tian H, Dye C, Lloyd-Smith JO, Eggo RM, Shrestha M, Scarpino SV, **Gutierrez B** *et al*. 2020. *Modelling COVID-19*. ***Nature Reviews Physics*** 2(6): 279-281.
- Rader B, Scarpino SV, Nande A, Hill AL, Adlam B, Reiner RC, Pigott DM, **Gutierrez B** *et al*. 2020. *Crowding and the shape of COVID-19 epidemics*. ***Nature Medicine*** 26: 1829-1834.
- de Souza WM, Buss LF, Candido DS, Carrera JP, Li S, Zarebski A, Moraes Pereira RH, Prete Jr CA, de Souza-Santos AA, Parag KV, Belotti CTD, Vincenti-Gonzalez M, Messina J, Sales FCS, Andrade PS, Nascimento VH, Ghilardi F, Abade L, **Gutierrez B** *et al*. 2020. *Epidemiological and clinical characteristics of the early phase of the COVID-19 epidemic in Brazil*. ***Nature Human Behaviour*** 4: 856-865.
- Kraemer MUG, Yang CH, **Gutierrez B**, Wu CH, Klein B, Pigott DM et al. 2020. The effect of human mobility and measures on the COVID-19 epidemic in China. **Science** 368(6490): 493-497.
- Xu B, **Gutierrez B**, Mekaru S, Sewalk K, Goodwin L, Loskill A, Cohn E, Hswen Y *et al*. 2020. *Epidemiological data from the COVID-19 outbreak, real-time case information*. ***Scientific Data*** 7(1): 1-6.

Work related to other genomic analyses of SARS-CoV-2.

- Kraemer MUG, Hill V, Ruis C, Dellicour S, Bajaj S, McCrone JT, Baele G, Parag KV, Battle AL, **Gutierrez B** *et al*. 2021. *Spatiotemporal invasion dynamics of SARS-CoV-2lineage B.1.1.7 emergence*. ***Science*** 373: 889-895.

Other papers not relevant to this thesis

- Escalera-Zamudio M, Golden M, **Gutiérrez B**, Thézé J, Keown JR, Carrique L et al. 2020. Parallel evolution in the emergence of highly pathogenic avian influenza A viruses. **Nature Communications** 11: 5511.
- Kraemer MUG, Pigott DM, Hill SC, Vanderslott S, Reiner RC, Stasse S, Brownstein JS, **Gutierrez B** *et al*. 2020. *Dynamics of conflict during the Ebola outbreak in the Democratic Republic of the Congo 2018–2019*. ***BMC Medicine*** 18: 1-10.
- Avanzato VA, Oguntuyo KY, Escalera-Zamudio M, **Gutierrez B**, Golden M, Kosavsky Pond SL et al. 2019. A structural basis for antibody-mediated neutralization of Nipah virus reveals a site of vulnerability at the fusion glycoprotein apex. **Proceedings of the National Academy of Sciences** 116(50

Appendix 2 Supplementary material for Chapter 2

The evolutionary dynamics of Oropouche Virus (OROV) in South America

Bernardo Gutierrez, Emma Wise, Steven Pullan, Christopher Logue, Thomas A. Bowden, Marina Escalera-Zamudio, Gabriel Trueba, Marcio Nunes, Nuno R. Faria, Oliver G. Pybus

Supplementary Figures

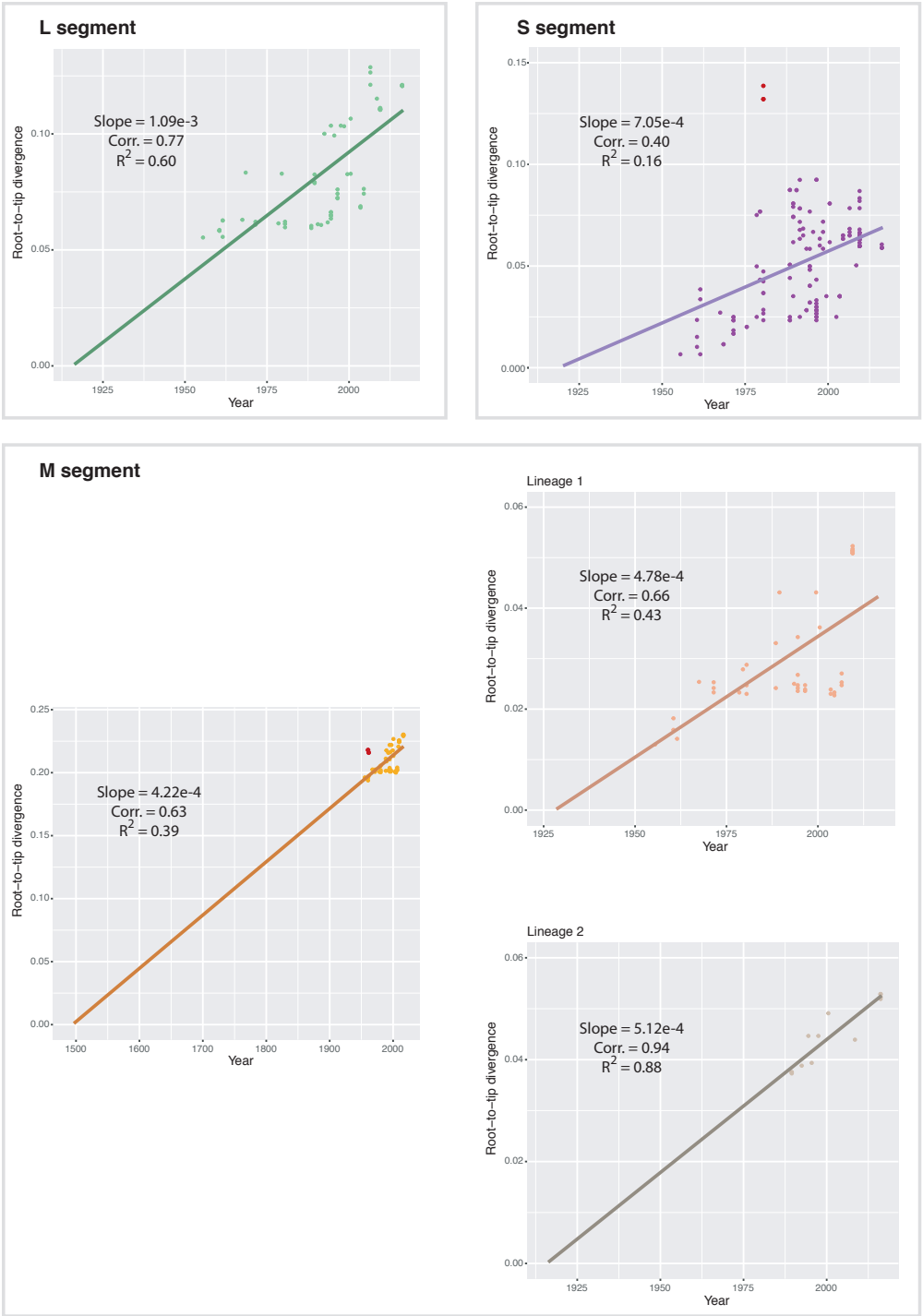
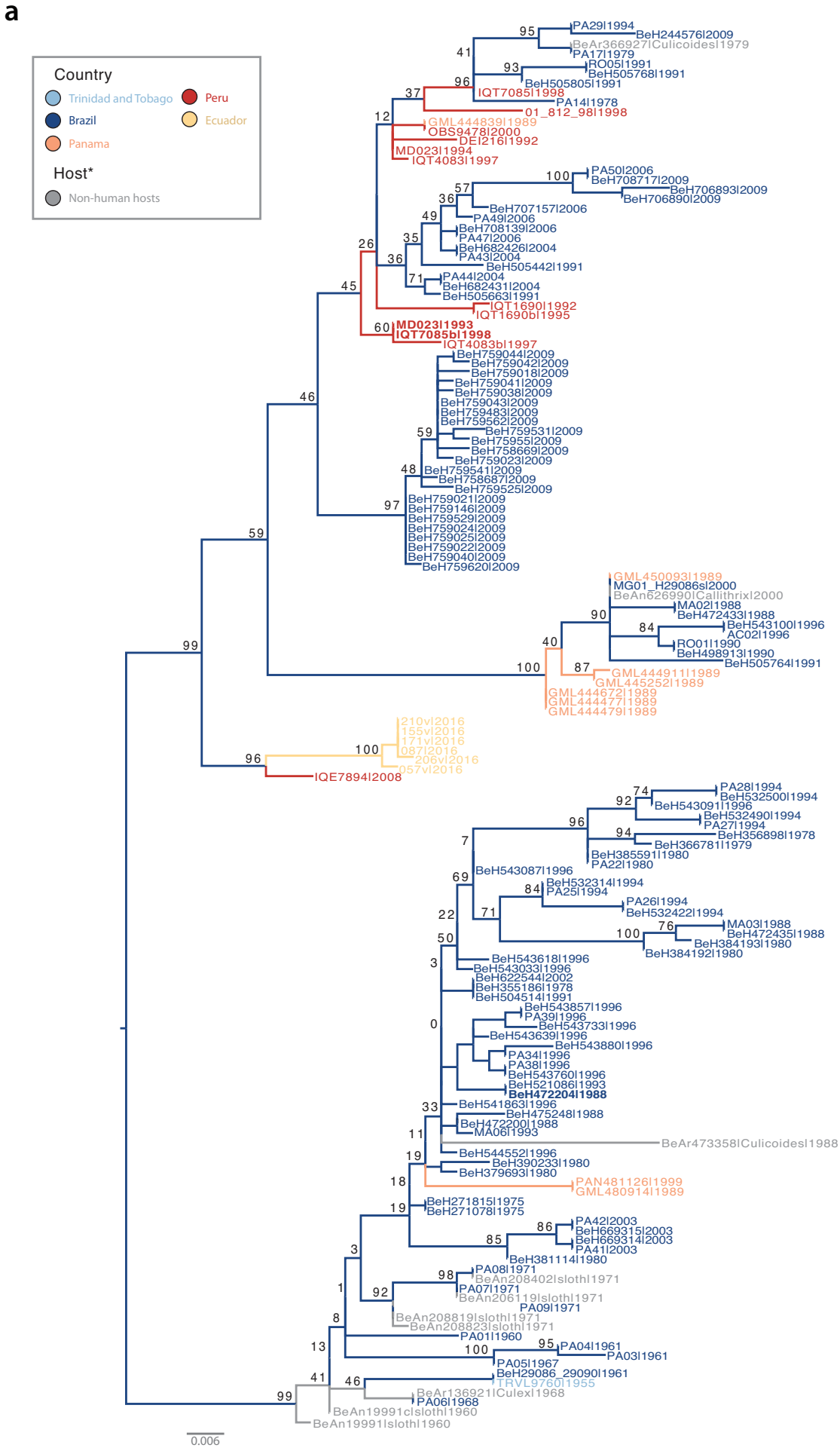


Figure S1 (above). Correlation plots between tip-to-root distances versus sampling estimated in TempEst for the three OROV genome segments. Outliers (determined through visual inspection) are highlighted in red. Sequences corresponding to the points highlighted in red (sequences PA01 and PA03 for the M segment; sequences AM01, AM03 and BeH389865 for the S segment) were excluded as outliers prior to the estimation of the individual plots for each M lineage and all further molecular clock analyses. Regression slope (interpreted as rough estimates of the mean evolutionary rate), correlation coefficients and R^2 values are shown for each plot.


a
Country
Trinidad and Tobago
Brazil
Panama
Peru
Ecuador
Host*
Non-human hosts
PA29|1994
BeH244576|2009
BeAr366927|Culicoides|1979
PA17|1979
RO05|1991
BeH505768|1991
BeH505805|1991
IQT7085|1998
PA14|1978
01_812_98|1998
GML444839|1989
OBS9478|2000
DEI216|1992
MD023|1994
IQT4083|1997
PA50|2006
BeH708717|2009
BeH706893|2009
BeH706890|2009
BeH707157|2006
PA49|2006
BeH708139|2006
PA47|2006
BeH682426|2004
PA43|2004
BeH505442|1991
PA44|2004
BeH682431|2004
BeH505663|1991
IQT1690|1992
IQT1690bl1995
MD023|1993
IQT7085bl1998
IQT4083bl1997
BeH759044|2009
BeH759042|2009
BeH759018|2009
BeH759041|2009
BeH759038|2009
BeH759043|2009
BeH759483|2009
BeH759562|2009
BeH759531|2009
BeH75955|2009
BeH758669|2009
BeH759023|2009
BeH759541|2009
BeH758687|2009
BeH759525|2009
BeH75902|2009
BeH759146|2009
BeH759529|2009
BeH759024|2009
BeH759025|2009
BeH759022|2009
BeH759040|2009
BeH759620|2009
GML450093|1989
MG01_H29086sl2000
BeAn626990|Callithrix|2000
MA02|1988
BeH472433|1988
BeH543100|1996
AC02|1996
RO01|1990
BeH498913|1990
BeH505764|1991
GML444911|1989
GML445252|1989
GML444672|1989
GML444477|1989
GML444479|1989
210vl2016
155vl2016
171vl2016
087|2016
206vl2016
057vl2016
IQE7894|2008
PA28|1994
BeH532500|1994
BeH543091|1996
BeH532490|1994
PA27|1994
BeH356898|1978
BeH366781|1979
BeH385591|1980
PA22|1980
BeH543087|1996
BeH532314|1994
PA25|1994
PA26|1994
BeH532422|1994
MA03|1988
BeH472435|1988
BeH384193|1980
BeH384192|1980
BeH543618|1996
BeH543033|1996
BeH622544|2002
BeH355186|1978
BeH504514|1991
BeH543857|1996
PA39|1996
BeH543733|1996
BeH543639|1996
BeH543880|1996
PA34|1996
PA38|1996
BeH543760|1996
BeH521086|1993
BeH472204|1988
BeH541863|1996
BeH475248|1988
BeH472200|1988
MA06|1993
BeAr473358|Culicoides|1988
BeH544552|1996
BeH390233|1980
BeH379693|1980
PAN481126|1999
GML480914|1989
BeH271815|1975
BeH271078|1975
PA42|2003
BeH669315|2003
BeH669314|2003
PA41|2003
BeH381114|1980
PA08|1971
BeAn208402|sloth|1971
PA07|1971
BeAn206119|sloth|1971
PA09|1971
BeAn208819|sloth|1971
BeAn208823|sloth|1971
PA01|1960
PA04|1961
PA03|1961
PA05|1967
BeH29086_29090|1961
TRVL9760|1955
BeAr136921|Culex|1968
PA06|1968
BeAn19991c|sloth|1960
BeAn19991|sloth|1960
0.006

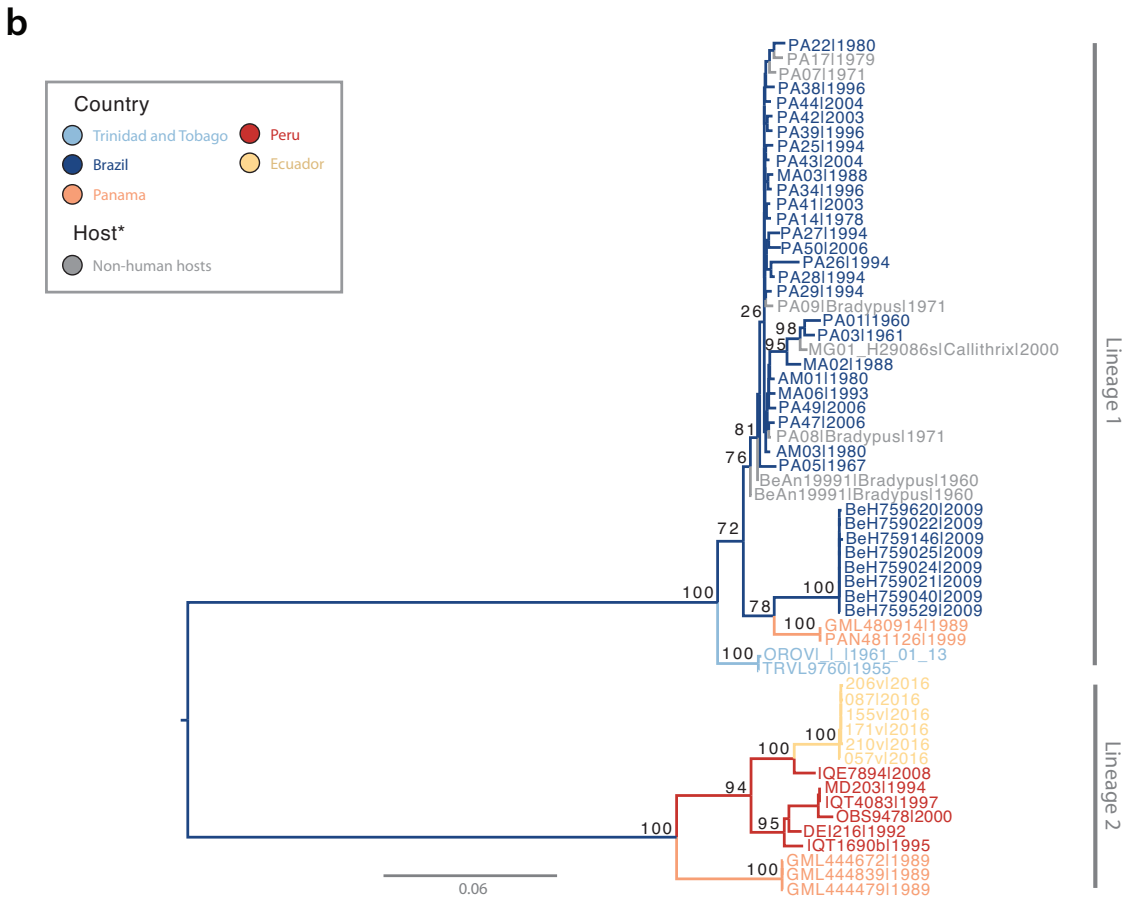
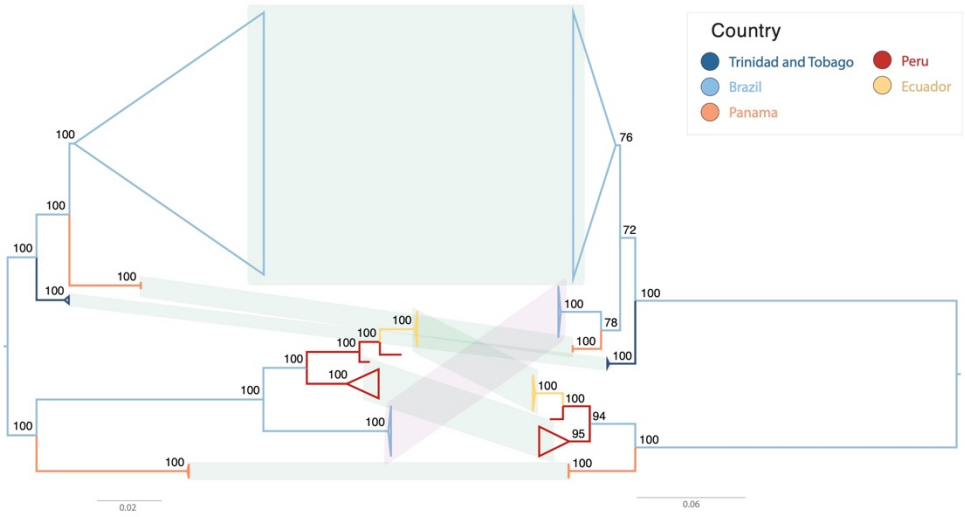
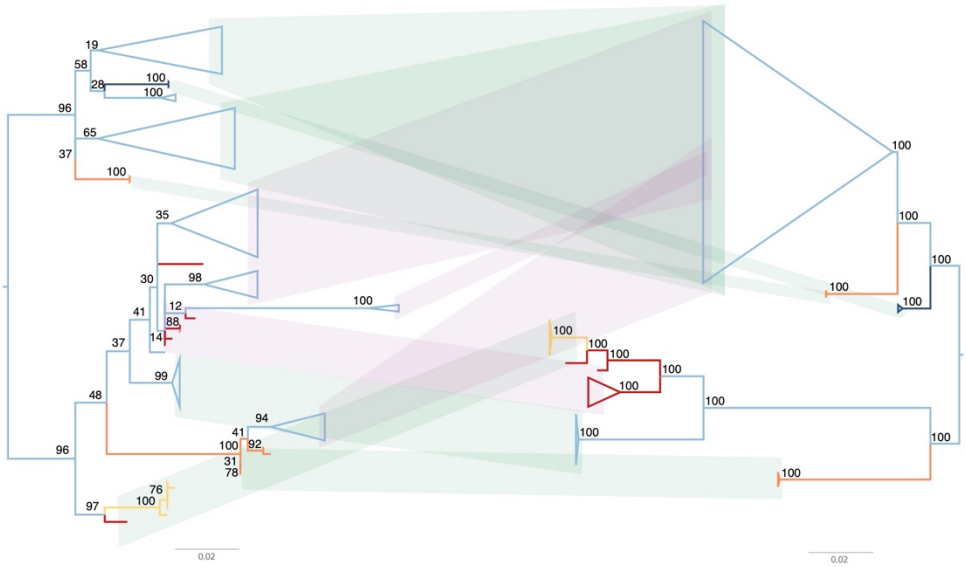

b
Country
Trinidad and Tobago
Brazil
Panama
Peru
Ecuador
Host*
Non-human hosts
PA22|1980
PA17|1979
PA07|1971
PA38|1996
PA44|2004
PA42|2003
PA39|1996
PA25|1994
PA43|2004
MA03|1988
PA34|1996
PA41|2003
PA14|1978
PA27|1994
PA50|2006
PA26|1994
PA28|1994
PA29|1994
PA09|Bradypus|1971
PA01|1960
PA03|1961
MG01_H29086s|Callithrix|2000
MA02|1988
AM01|1980
MA06|1993
PA49|2006
PA47|2006
PA08|Bradypus|1971
AM03|1980
PA05|1967
BeAn19991|Bradypus|1960
BeAn19991|Bradypus|1960
BeH759620|2009
BeH759022|2009
BeH759146|2009
BeH759025|2009
BeH759024|2009
BeH759021|2009
BeH759040|2009
BeH759529|2009
GML480914|1989
PAN481126|1999
OROVI_I_|1961_01_13
TRVL9760|1955
206v|2016
087|2016
155v|2016
171v|2016
210v|2016
057v|2016
IQE7894|2008
MD203|1994
IQT4083|1997
OBS9478|2000
DEI216|1992
IQT1690b|1995
GML444672|1989
GML444839|1989
GML444479|1989
Lineage 1
Lineage 2
0.06
c
PA06|Culex|1968
PA17|Culicoides|1979
MG01|Callithrix|2000
AC02|1996
PA39|1996
PA34|1996
PA38|1996
PA28|1994
PA25|1994
PA29|1994
PA26|1994
PA27|1994
MA06|1993
RO05|1991
RO01|1990
PA22|1980
AM03|1980
PA50|2006
PA47|2006
PA49|2006
PA43|2004
PA44|2004
PA41|2003
PA42|2003
PA14|1978
MA02|1988
PA08|Bradypus|1971
PA07|Bradypus|1971
PA09|Bradypus|1971
PA03|1961
PA04|1961
PA05|1967
AM01|1980
MA03|1988
PA01|1960
BeAn19991|Bradypus|1960
BeAn19991b|Bradypus|1960
PAN481126|1999
GML480914|1989
OROVILI_|Trinidad|1961
TRVL9760|1955
GML444479|1989
GML444672|1989
GML444839|1989
057v|2016
155v|2016
206v|2016
210v|2016
171v|2016
087|2016
IQE7894|2008
IQT7085|1998
DEI216|1992
IQT4083|1997
MD203|1994
OBS9478|2000
IQT1690b|1995
BeH759620|2009
BeH759146|2009
BeH759529|2009
BeH759022|2009
BeH759025|2009
BeH759024|2009
BeH759040|2009
BeH759021|2009
0.02

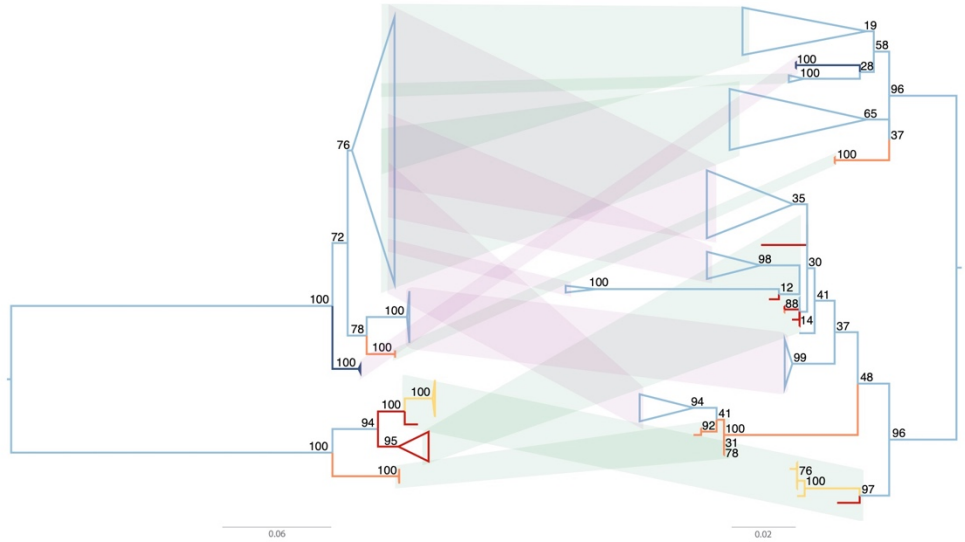
Figure S2 (above). Maximum Likelihood trees for the three OROV genome segments. (**a**) Small (S) segment phylogeny. (**b**) Medium (M) segment phylogeny. (**c**) Large (L) segment phylogeny. Tip labels and corresponding branches are colour coded by the country of origin for each sample, with grey samples indicating sequences obtained from non-human hosts and vectors (species indicated in the tip label). Node support from 100 bootstrap replicates is shown for the basal internal nodes of each tree.



L Segment M Segment



S Segment L Segment



M Segment S Segment

Figure S3 (above). Pairwise comparison of maximum-likelihood phylogenetic trees of the three OROV genome segments suggest widespread reassortment. The L and S trees are midpoint rooted, and the M tree was re-rooted to maximise its topological congruency with its closest match, the L tree. Node support representing 100 bootstraps is shown on the tree nodes, where important monophyletic clades are collapsed in all trees (most collapsed nodes have high node support). All branches are coloured according to the most probable location using on a parsimony approach; colours match those used Fig. 1. Topologically similar clusters are highlighted in green, while potentially reassortant clusters are highlighted in purple.

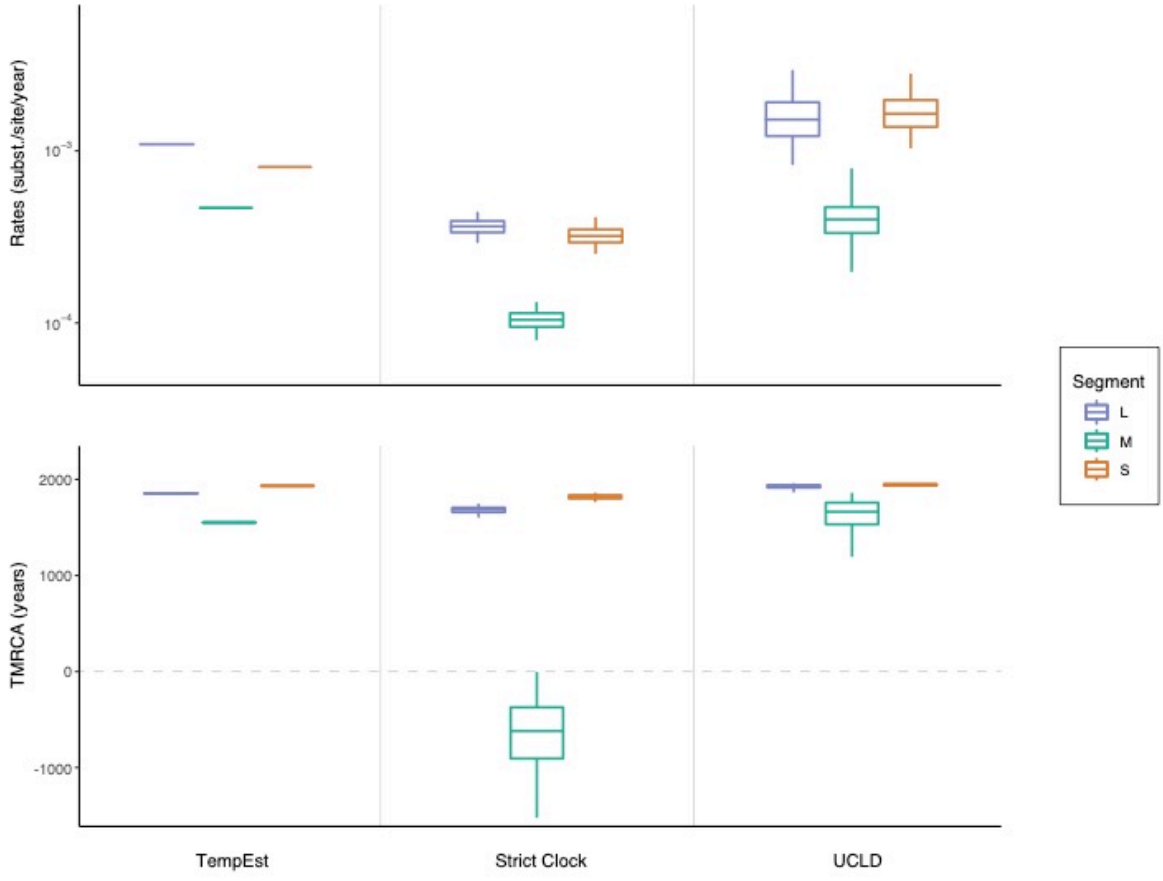


Figure S4 (above). Comparison of evolutionary rate estimate (upper panel) and times to the most recent common ancestor estimates (lower panel) for the three OROV genome segments estimated through different methods and clock models: a root-to-tip regression slope (TempEst), and a phylogenetically co-inferred rate using a strict clock model and a relaxed clock model (UCLD).

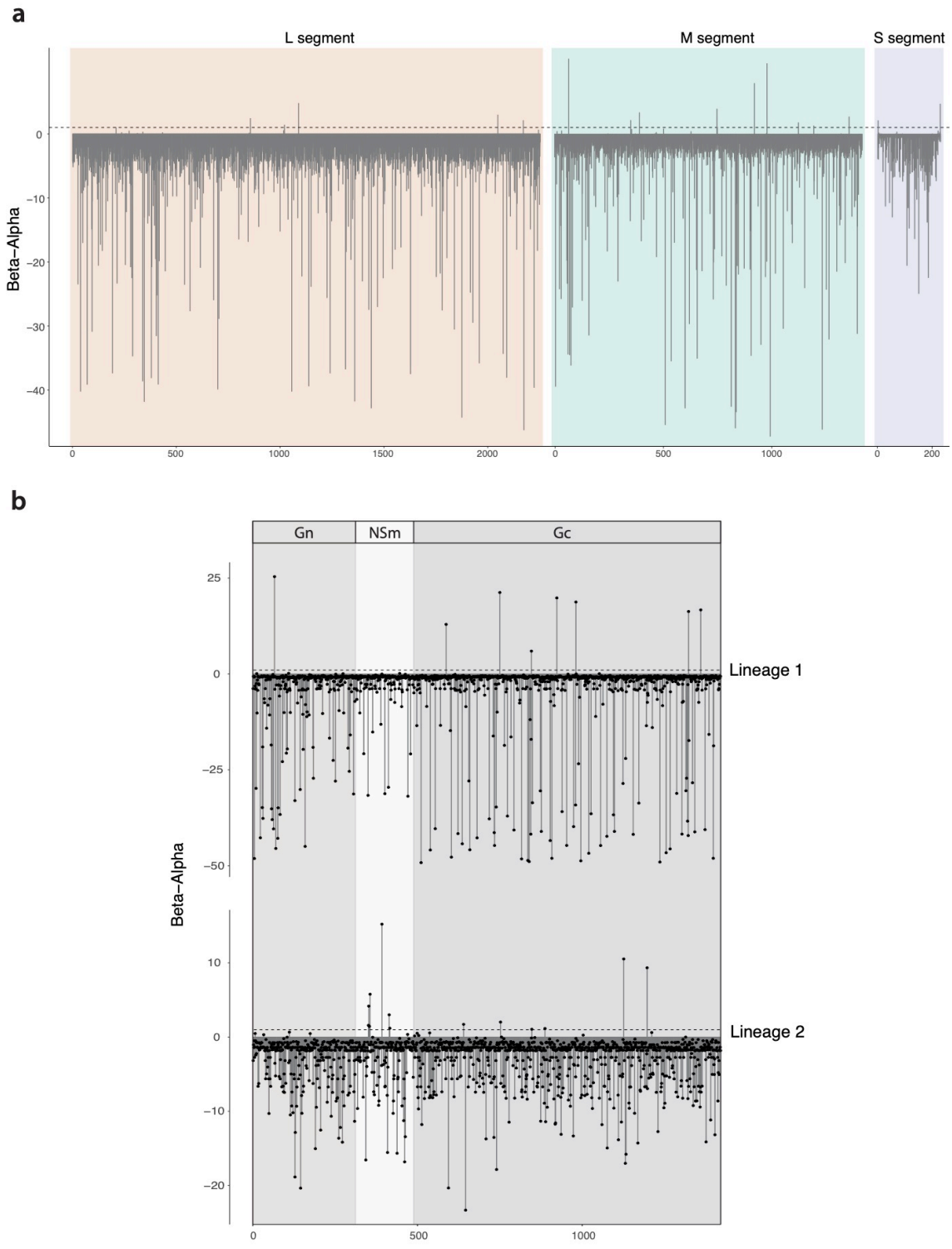


Figure S5. Site-wise differences between the non-synonymous (β) and synonymous (α) substitution rates in the OROV genome. The three genome segments (**a**) show a limited number of sites that evolve at a higher non-synonymous substitution rate (>1, marked by the

dotted line), suggestive of adaptive selection. Within the M segment (**b**), different codons (mostly on different protein reading frames) show evidence for adaptive selection.

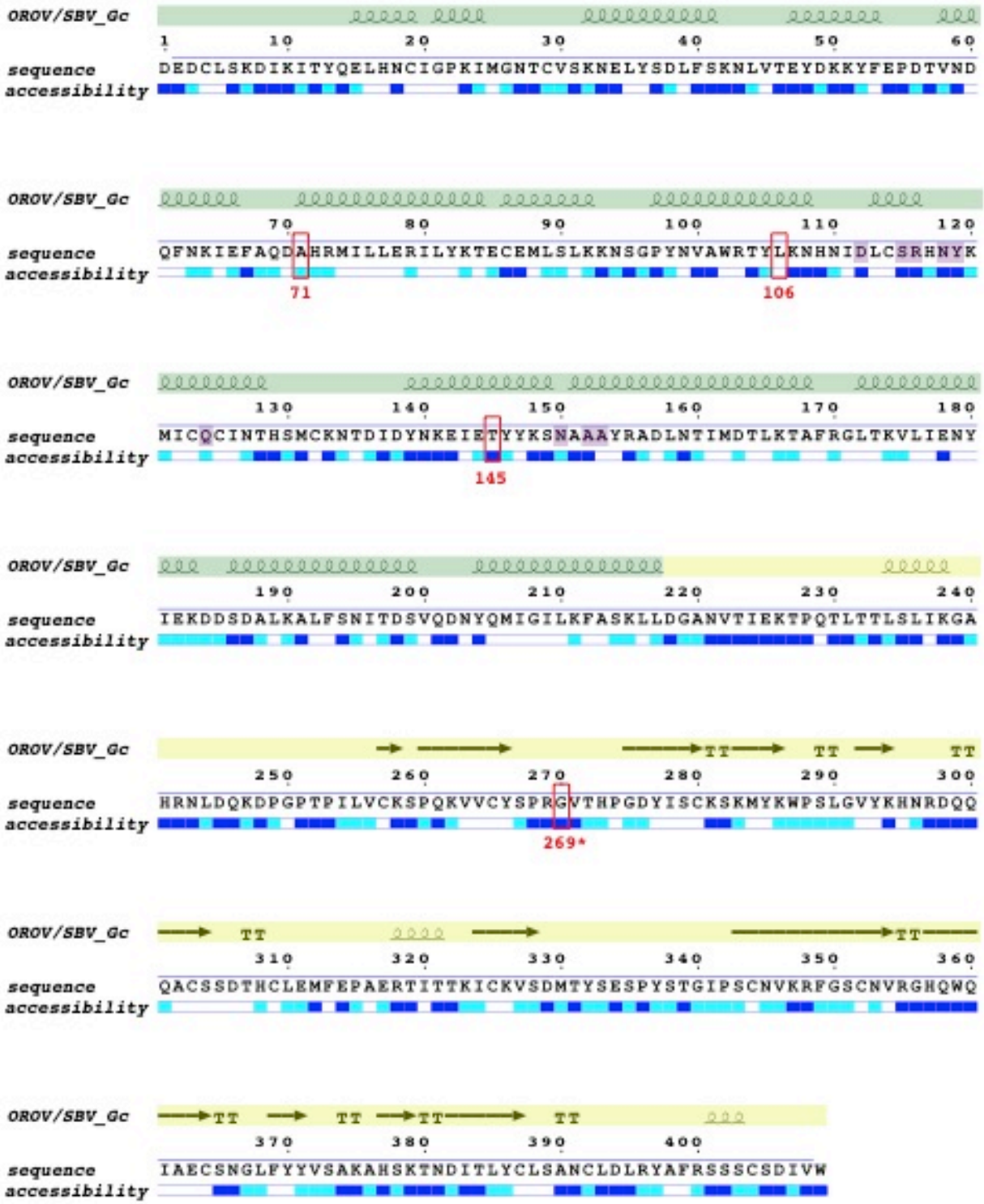


Figure S6 (above). Relative residue accessibility for the OROV/SBV composite model of the Gc protein, based on its atomic structure and amino acid sequence, using the ESPript server. Accessibility is scored as buried residues (white), intermediate residues (cyan) and exposed residues (blue). Red boxes show sites identified to be evolving under positive selection, with the red numbering highlighting the equivalent numbering at the OROV Gc sequence. An asterisk represents the site mapped onto the SBV component of the composite model, hence the numbering discrepancy. Secondary structures are shown above the sequence and highlighted based on the virus from which each component of the sequence was used to construct the composite model (green for OROV, yellow for SBV). Residues that directly play a role in the trimerisation interface of the head spike architecture, as identified in the PDBePISA server, are highlighted in purple.

Supplementary Tables

Table S1. Summary of all isolates included in this study, with accession numbers for each genome segment (including six new sequences from Ecuador).

ID	Host	City	Province/State	Country	Year	S	M	L
AC02	Homo sapiens	Xapuri	Acre	Brazil	1996	MG747503	MG747504	MG747505
AM01	Homo sapiens	Manaus	Amazonas (BR)	Brazil	1980	MG747506	MG747507	MG747508
AM03	Homo sapiens	Manaus	Amazonas (BR)	Brazil	1980	MG747509	MG747510	MG747511
MA02	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	MG747512	MG747513	MG747514
MA03	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	MG747515	MG747516	MG747517
MA06	Homo sapiens	Barra da Corda	Maranhao	Brazil	1993	MG747518	MG747519	MG747520
MG01	Callithrix sp.	Arino	Minas Gerais	Brazil	2000	MG747521	MG747522	MG747523
01-812-98	Homo sapiens	Iquitos	Loreto	Peru	1998	AF164553	---	---
D-057/057v	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MK506818	MK506823	MK506828
D-087/087	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MF926352	MF926353	MF926354
D-155/155v	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MK506819	MK506824	MK506829
D-171/171v	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MK506820	MK506825	MK506830
D-206/206v	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MK506821	MK506826	MK506831
D-210/210v	Homo sapiens	Esmeraldas	Esmeraldas	Ecuador	2016	MK506822	MK506827	MK506832
BeAn19991	Bradypus trydactylus	Sao Miguel do Guama	Para	Brazil	1960	KP052852	KP052851	KP052850
BeAn19991	Bradypus trydactylus	Sao Miguel do Guama	Para	Brazil	1960	AF164532	AF441119	KP052850
BeAn206119	Bradypus trydactylus	Maracana	Para	Brazil	1971	HQ830378	---	---
BeAn208402	Bradypus trydactylus	Maracana	Para	Brazil	1971	HQ830379	---	---
BeAn208819	Bradypus trydactylus	Maracana	Para	Brazil	1971	HQ830380	---	---
BeAn208823	Bradypus trydactylus	Maracana	Para	Brazil	1971	AY993912	---	---
BeAn626990	Callithrix sp.	Arino	Minas Gerais	Brazil	2000	AY117135	---	---
BeAr136921	Culex quinquefasciatus	Belem	Para	Brazil	1968	HM470111	---	---
BeAr366927	Culicoides paraensis	Belem	Para	Brazil	1979	HQ830448	---	---
BeAr473358	Culicoides paraensis	Porto Franco	Maranhao	Brazil	1988	AF164539	---	---
BeH244576	Homo sapiens	Belem	Para	Brazil	2009	HQ830444	---	---
BeH271078	Homo sapiens	Santarem	Para	Brazil	1975	HQ830445	---	---
BeH271815	Homo sapiens	Santarem	Para	Brazil	1975	AF164533	---	---
BeH29086	Homo sapiens	Belem	Para	Brazil	1961	HM470108	---	---
BeH355186	Homo sapiens	Tome-Acu	Para	Brazil	1978	HQ830446	---	---
BeH356898	Homo sapiens	Belem	Para	Brazil	1978	HQ830447	---	---
BeH366781	Homo sapiens	Belem	Para	Brazil	1979	HQ830449	---	---
BeH379693	Homo sapiens	Castanhal	Para	Brazil	1980	AF164534	---	---

Table S1. (Continued)

ID	Host	City	Province/State	Country	Year	S	M	L
BeH381114	Homo sapiens	Belem	Para	Brazil	1980	AF164535	---	---
BeH384192	Homo sapiens	Portel	Para	Brazil	1980	HQ830450	---	---
BeH384193	Homo sapiens	Portel	Para	Brazil	1980	HQ830451	---	---
BeH385591	Homo sapiens	Belem	Para	Brazil	1980	HQ830452	---	---
BeH389865	Homo sapiens	Manaus	Amazonas (BR)	Brazil	1980	HQ830453	---	---
BeH390233	Homo sapiens	Manaus	Amazonas (BR)	Brazil	1980	AF164536	---	---
BeH472200	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	AF164537	---	---
BeH472204	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	AF164538	---	---
BeH472433	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	HQ830455	---	---
BeH472435	Homo sapiens	Porto Franco	Maranhao	Brazil	1988	HQ830456	---	---
BeH475248	Homo sapiens	Tucurui	Para	Brazil	1988	AF164540	---	---
BeH498913	Homo sapiens	Machadinho D'Oeste	Rondonia	Brazil	1990	HQ830457	---	---
BeH504514	Homo sapiens	Americano	Para	Brazil	1991	AF164541	---	---
BeH505442	Homo sapiens	Ouro Preto do Oeste	Rondonia	Brazil	1991	AF164542	---	---
BeH505663	Homo sapiens	Ariquemes	Rondonia	Brazil	1991	AF164543	---	---
BeH505764	Homo sapiens	Ariquemes	Rondonia	Brazil	1991	HQ830458	---	---
BeH505768	Homo sapiens	Ariquemes	Rondonia	Brazil	1991	HQ830459	---	---
BeH505805	Homo sapiens	Ariquemes	Rondonia	Brazil	1991	HQ830460	---	---
BeH521086	Homo sapiens	Barra da Corda	Maranhao	Brazil	1993	AY704559	---	---
BeH532314	Homo sapiens	Serra Pelada	Para	Brazil	1994	HQ830461	---	---
BeH532422	Homo sapiens	Serra Pelada	Para	Brazil	1994	HQ830462	---	---
BeH532490	Homo sapiens	Serra Pelada	Para	Brazil	1994	HQ830463	---	---
BeH532500	Homo sapiens	Serra Pelada	Para	Brazil	1994	HQ830464	---	---
BeH541863	Homo sapiens	Altamira	Para	Brazil	1996	AF164544	---	---
BeH543033	Homo sapiens	Oriximina	Para	Brazil	1996	AF164545	---	---
BeH543087	Homo sapiens	Xapuri	Acre	Brazil	1996	AF164547	---	---
BeH543091	Homo sapiens	Xapuri	Acre	Brazil	1996	HQ830465	---	---
BeH543100	Homo sapiens	Xapuri	Acre	Brazil	1996	HQ830466	---	---
BeH543618	Homo sapiens	Oriximina	Para	Brazil	1996	AF164548	---	---
BeH543639	Homo sapiens	Oriximina	Para	Brazil	1996	AY704562	---	---
BeH543733	Homo sapiens	Oriximina	Para	Brazil	1996	AY704560	---	---
BeH543760	Homo sapiens	Oriximina	Para	Brazil	1996	AY704568	---	---
BeH543857	Homo sapiens	Oriximina	Para	Brazil	1996	AY704566	---	---
BeH543880	Homo sapiens	Oriximina	Para	Brazil	1996	AY704565	---	---
BeH544552	Homo sapiens	Altamira	Para	Brazil	1996	AF164546	---	---
BeH622544	Homo sapiens	Parana	Tocantins	Brazil	2002	EF467368	---	---
BeH669314	Homo sapiens	Parauapebas	Para	Brazil	2003	EF467370	---	---
BeH669315	Homo sapiens	Parauapebas	Para	Brazil	2003	EF467369	---	---

Table S1. (Continued)

ID	Host	City	Province/State	Country	Year	S	M	L
BeH682426	Homo sapiens	Porto de Moz	Para	Brazil	2004	EF467371	---	---
BeH682431	Homo sapiens	Porto de Moz	Para	Brazil	2004	EF467372	---	---
BeH706890	Homo sapiens	Igarape Acu	Para	Brazil	2009	HQ830473	---	---
BeH706893	Homo sapiens	Igarape Acu	Para	Brazil	2009	HQ830474	---	---
BeH707157	Homo sapiens	Maracana	Para	Brazil	2006	HQ830476	---	---
BeH708139	Homo sapiens	Magalhaes Barata	Para	Brazil	2006	HQ830475	---	---
BeH708717	Homo sapiens	Maracana	Para	Brazil	2009	HQ830477	---	---
BeH758669	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830479	---	---
BeH758687	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830478	---	---
BeH759018	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830486	---	---
BeH759021	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691608	KP691607	KP691606
BeH759022	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691611	KP691610	KP691609
BeH759023	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830487	---	---
BeH759024	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691605	KP691604	KP691603
BeH759025	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691614	KP691613	KP691612
BeH759038	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830484	---	---
BeH759040	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691617	KP691616	KP691615
BeH759041	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830488	---	---
BeH759042	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830489	---	---
BeH759043	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830490	---	---
BeH759044	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830491	---	---
BeH759146	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691632	---	---
BeH759483	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830492	---	---
BeH759525	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830480	---	---
BeH759529	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691620	KP691619	KP691618
BeH759531	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830482	---	---
BeH759541	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830481	---	---
BeH759558	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830483	---	---
BeH759562	Homo sapiens	Mazagao	Amapa	Brazil	2009	HQ830485	---	---
BeH759620	Homo sapiens	Mazagao	Amapa	Brazil	2009	KP691623	KP691622	KP691621
DEI216	Homo sapiens	Iquitos	Loreto	Peru	1992	KP795074	KP795073	KP795072
GML444477	Homo sapiens	Chame	Panama	Panama	1989	AF164555	---	---
GML444479	NA	NA	Panama	Panama	1989	KC759130	KC759129	KC759128
GML444672	Homo sapiens	San Miguelito	Panama	Panama	1989	KP795077	KP795076	KP795075
GML444839	Homo sapiens	Bejuco	Panama	Panama	1989	KP795080	KP795079	KP795078
GML444911	Homo sapiens	Chame	Panama	Panama	1989	AF164556	---	---
GML445252	Homo sapiens	San Miguelito	Panama	Panama	1989	AF164557	---	---

Table S1. (Continued)

ID	Host	City	Province/State	Country	Year	S	M	L
GML450093	Homo sapiens	Chilibre	Panama	Panama	1989	AF164558	---	---
GML480914	Homo sapiens	NA	Panama	Panama	1989	KP795083	KP795082	KP795081
IQE7894	Homo sapiens	Iquitos	Loreto	Peru	2008	KP795086	KP795085	KP795084
IQT1690	Homo sapiens	Iquitos	Loreto	Peru	1992	KC759127	KP795088	KC759125
IQT1690*	Homo sapiens	Iquitos	Loreto	Peru	1995	KP795089	---	KP795087
IQT4083	Homo sapiens	Iquitos	Loreto	Peru	1997	KP795092	KP795091	KP795090
IQT4083	Homo sapiens	Iquitos	Loreto	Peru	1997	AF164552	---	---
IQT7085	Homo sapiens	Iquitos	Loreto	Peru	1998	KP795095	---	KP795093
IQT7085	Homo sapiens	Iquitos	Loreto	Peru	1998	AF164554	---	---
MD023	Homo sapiens	Madre de Dios	Madre de Dios	Peru	1993	AF164550	KP795097	KP795096
OBS9478	Homo sapiens	Bagua	Amazonas (PE)	Peru	2000	KP795101	KP795100	KP795099
PAN481126	Homo sapiens	Colon	Colon	Panama	1999	KP795104	KP795103	KP795102
TRVL9760	Homo sapiens	Vega de Oropouche	Trinidad	Trinidad	1955	KP026181	KP026180	KP026179
PA01	Ochlerotatus serratus	Ipixuna	Para	Brazil	1960	MG747524	MG7475215	MG747526
PA03	Homo sapiens	Belem	Para	Brazil	1961	MG747527	MG747528	MG747529
PA04	Homo sapiens	Belem	Para	Brazil	1961	MG747530	MG747531	MG747532
PA05	Homo sapiens	Braganca	Para	Brazil	1967	MG747533	MG747534	MG747535
PA06	Culex quinquefasciatus	Belem	Para	Brazil	1968	MG747536	MG747537	MG747538
PA07	Bradypus trydactylus	Maracana	Para	Brazil	1971	MG747539	MG747540	MG747541
PA08	Bradypus trydactylus	Maracana	Para	Brazil	1971	MG747542	MG747543	MG747544
PA09	Bradypus trydactylus	Maracana	Para	Brazil	1971	MG747545	MG747546	MG747547
PA14	Homo sapiens	Ananindeua	Para	Brazil	1978	MG747548	MG747549	MG747550
PA17	Culicoides paraensis	Belem	Para	Brazil	1979	MG747551	MG747552	MG747553
PA22	Homo sapiens	Belem	Para	Brazil	1980	MG747554	MG747555	MG747556
PA25	Homo sapiens	Serra Pelada	Para	Brazil	1994	MG747557	MG747558	MG747559
PA26	Homo sapiens	Serra Pelada	Para	Brazil	1994	MG747560	MG747561	MG747562
PA27	Homo sapiens	Serra Pelada	Para	Brazil	1994	MG747563	MG747564	MG747565
PA28	Homo sapiens	Serra Pelada	Para	Brazil	1994	MG747566	MG747567	MG747568
PA29	Homo sapiens	Altamira	Para	Brazil	1994	MG747569	MG747570	MG747571
PA34	Homo sapiens	Oriximina	Para	Brazil	1996	MG747572	MG747573	MG747574
PA38	Homo sapiens	Oriximina	Para	Brazil	1996	MG747575	MG747576	MG747577
PA39	Homo sapiens	Oriximina	Para	Brazil	1996	MG747578	MG747579	MG747580
PA41	Homo sapiens	Parauapebas	Para	Brazil	2003	MG747581	MG747582	MG747583
PA42	Homo sapiens	Parauapebas	Para	Brazil	2003	MG747584	MG747585	MG747586
PA43	Homo sapiens	Porto de Moz	Para	Brazil	2004	MG747587	MG747588	MG747589

Table S1. (Continued)

ID	Host	City	Province/State	Country	Year	S	M	L
PA43	Homo sapiens	Porto de Moz	Para	Brazil	2004	MG747587	MG747588	MG747589
PA44	Homo sapiens	Porto de Moz	Para	Brazil	2004	MG747590	MG747591	MG747592
PA47	Homo sapiens	Magalhaes Barata	Para	Brazil	2006	MG747593	MG747594	MG747595
PA49	Homo sapiens	Magalhaes Barata	Para	Brazil	2006	MG747596	MG747597	MG747598
PA50	Homo sapiens	Maracana	Para	Brazil	2006	MG747599	MG747600	MG747601
RO01	Homo sapiens	Machadinho D'Oeste	Rondonia	Brazil	1990	MG747602	MG747603	MG747604
RO05	Homo sapiens	Ariquemes	Rondonia	Brazil	1991	MG747605	MG747606	MG747607
Trinidad1961	Homo sapiens	Vega de Oropouche	Trinidad	Trinidad	1961	---	MF620128	MF620129

*June 1995 sequence excluded from analysis.

Table S2. Links to the data exploration tool Microreact for the three OROV genome segments analysed in this study.

OROV segment	Microreact link
L	https://microreact.org/project/OROV_America_L?lu=y
M	https://microreact.org/project/OROV_America_M?lu=y
S	https://microreact.org/project/OROV_America_S?lu=y

Table S3. Allele frequencies for codons under adaptive selection in the OROV genome.

Protein	Codon	Identity	Frequency
Gn	66	K	0.444
		R	0.333
		A	0.190
		T	0.032
NSm	86	R	0.651
		N	0.206
		S	0.143
Gc	269	D	0.656
		G	0.344
	442	T	0.734
		S	0.188
		I	0.078

Appendix 3 Supplementary material for Chapter 3

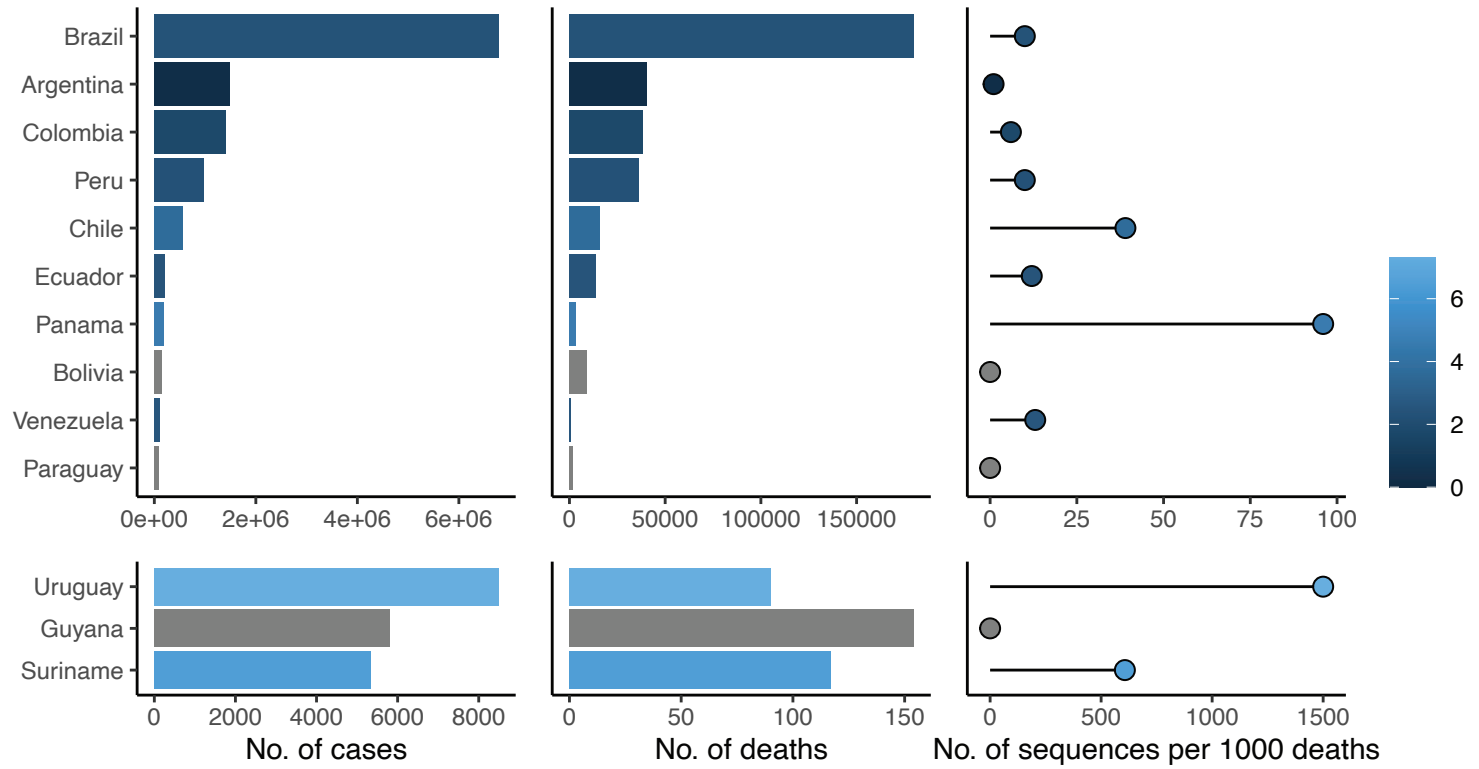


Fig S1. Comparison of the reported numbers of cases and deaths and the publicly available SARS-CoV-2 genomes generated per country in South America. Numbers of sequences were taken from the GISAID database, and the reported numbers of cases and deaths were taken from Our World in Data. Data retrieved on February 14 2021, with a cutoff date on 2020-12-10. Lower panels show data from countries at a different scale compared to the upper panels for visualisation purposes. Color scheme shows the (logarithmic) numbers of sequences per cumulative number of reported death.

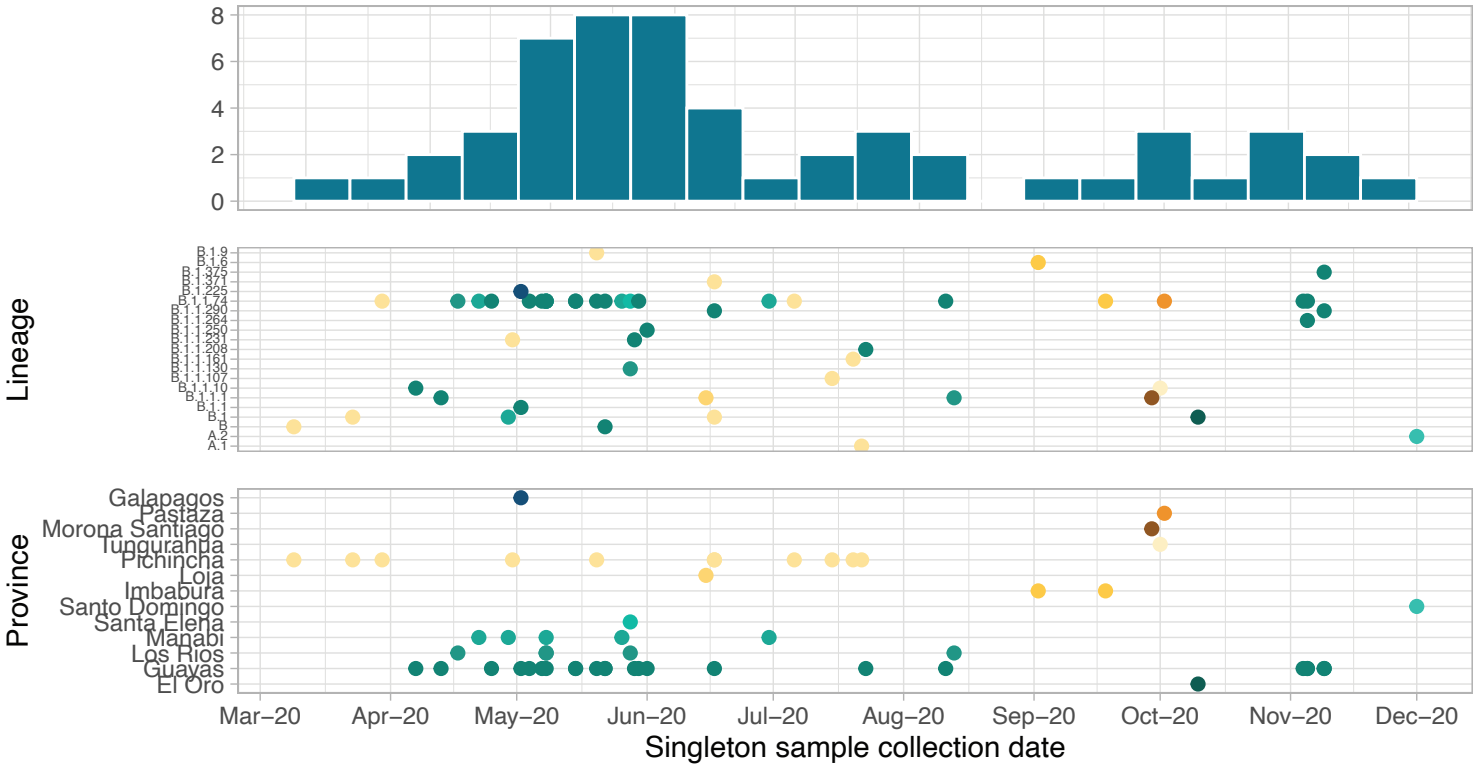


Fig S2. Summary of singletons identified in Ecuador. The upper panel shows the number of singletons per two-week epidemiological weeks. The middle panel shows the Pango lineages to which singletons were assigned, and the lower panel shows the province where singletons were identified by sample collection date. Dot colours correspond to the province where the samples were collected.

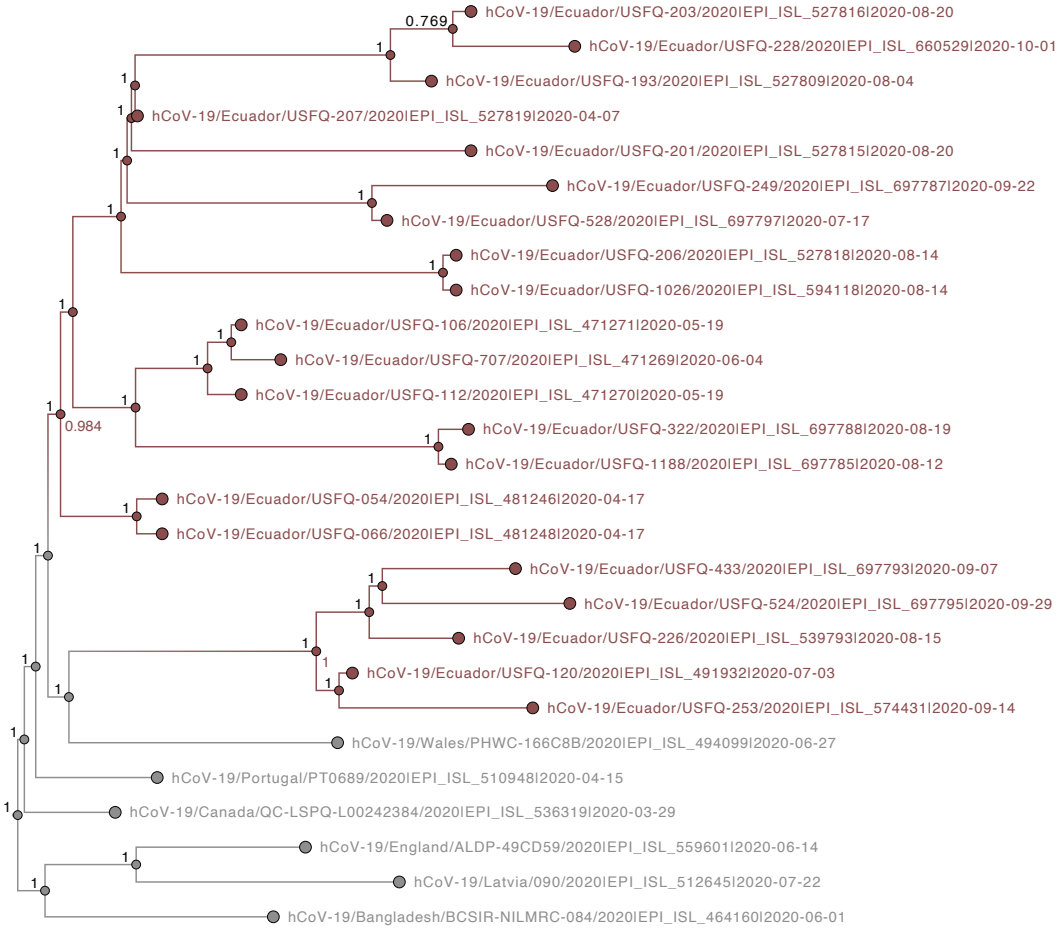


Fig S3. Subtree from the MCC tree, corresponding to transmission lineages group D. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

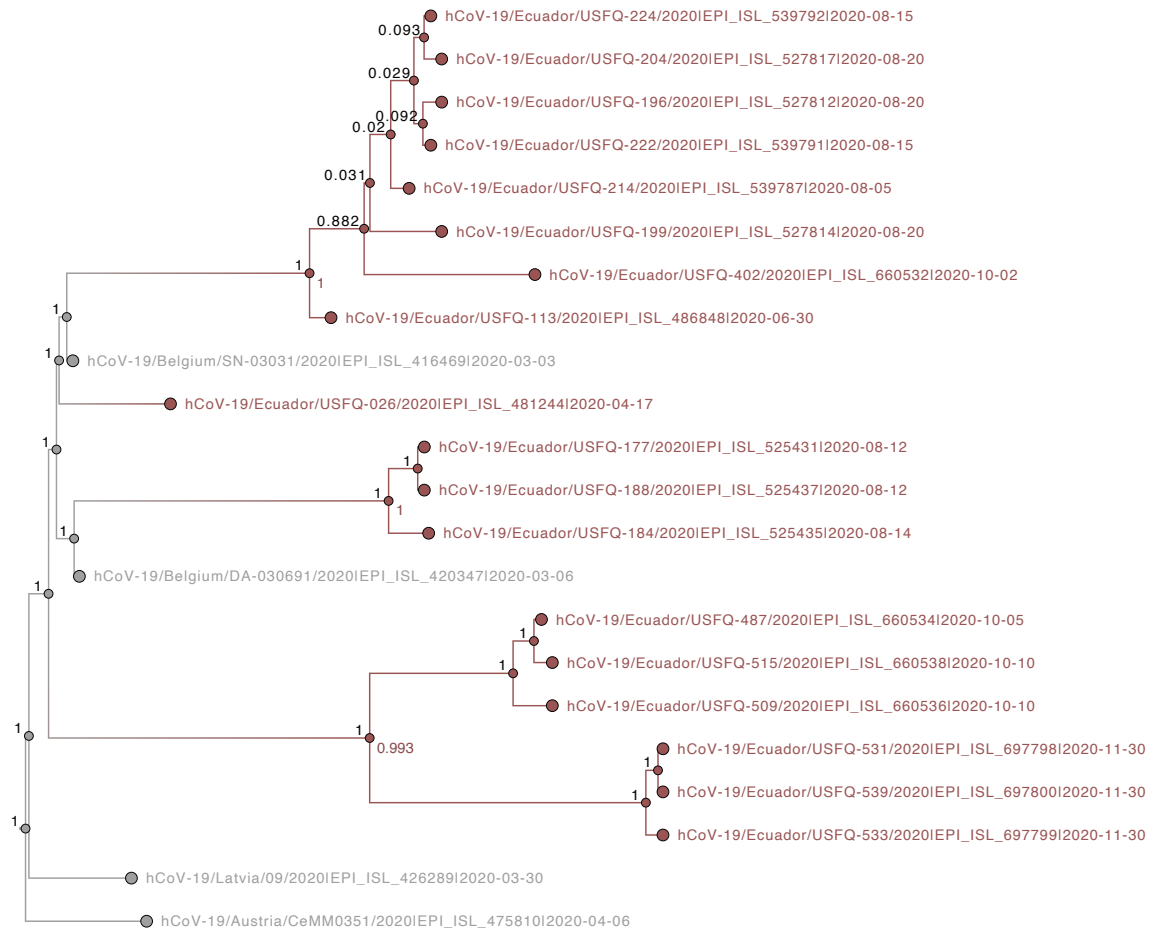


Fig S4. Subtree from the MCC tree, corresponding to transmission lineages group H. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

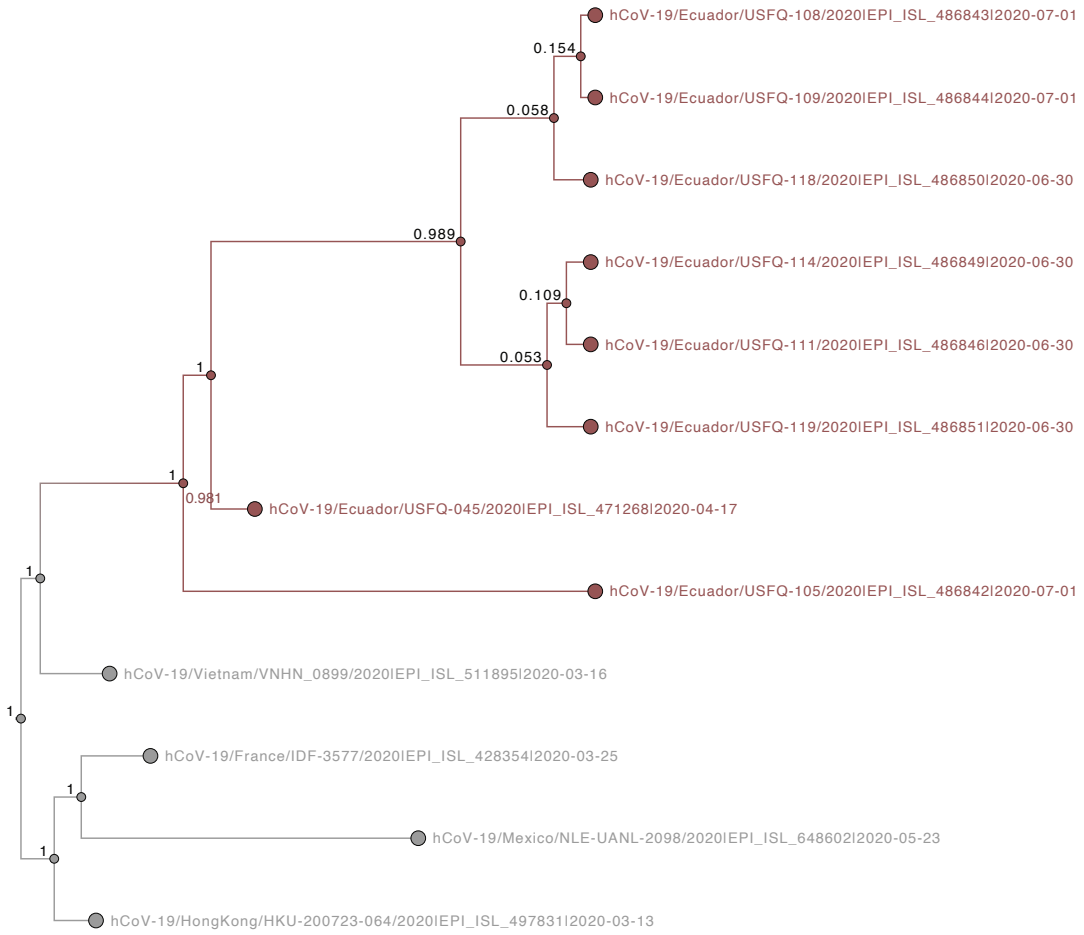


Fig S5. Subtree from the MCC tree, corresponding to transmission lineage G. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

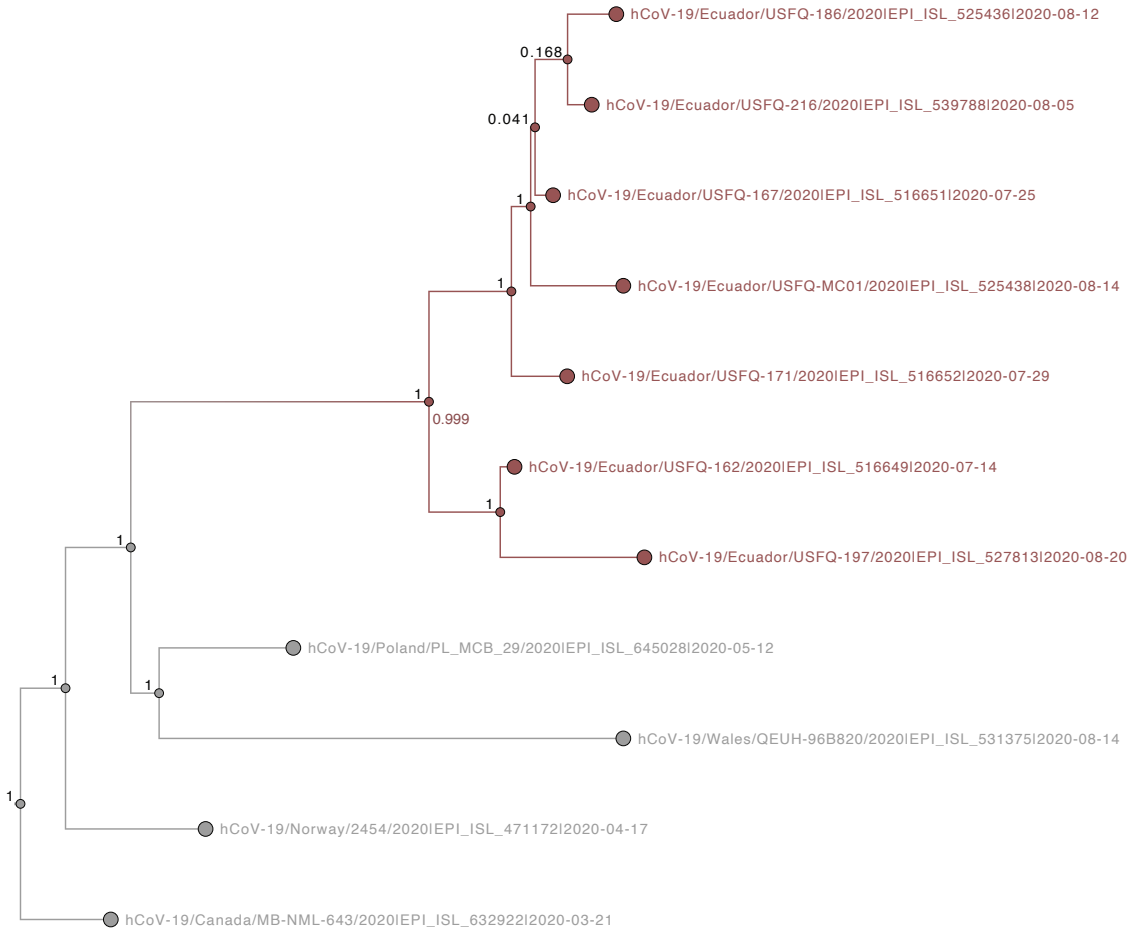


Fig S6. Subtree from the MCC tree, corresponding to transmission lineage O. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

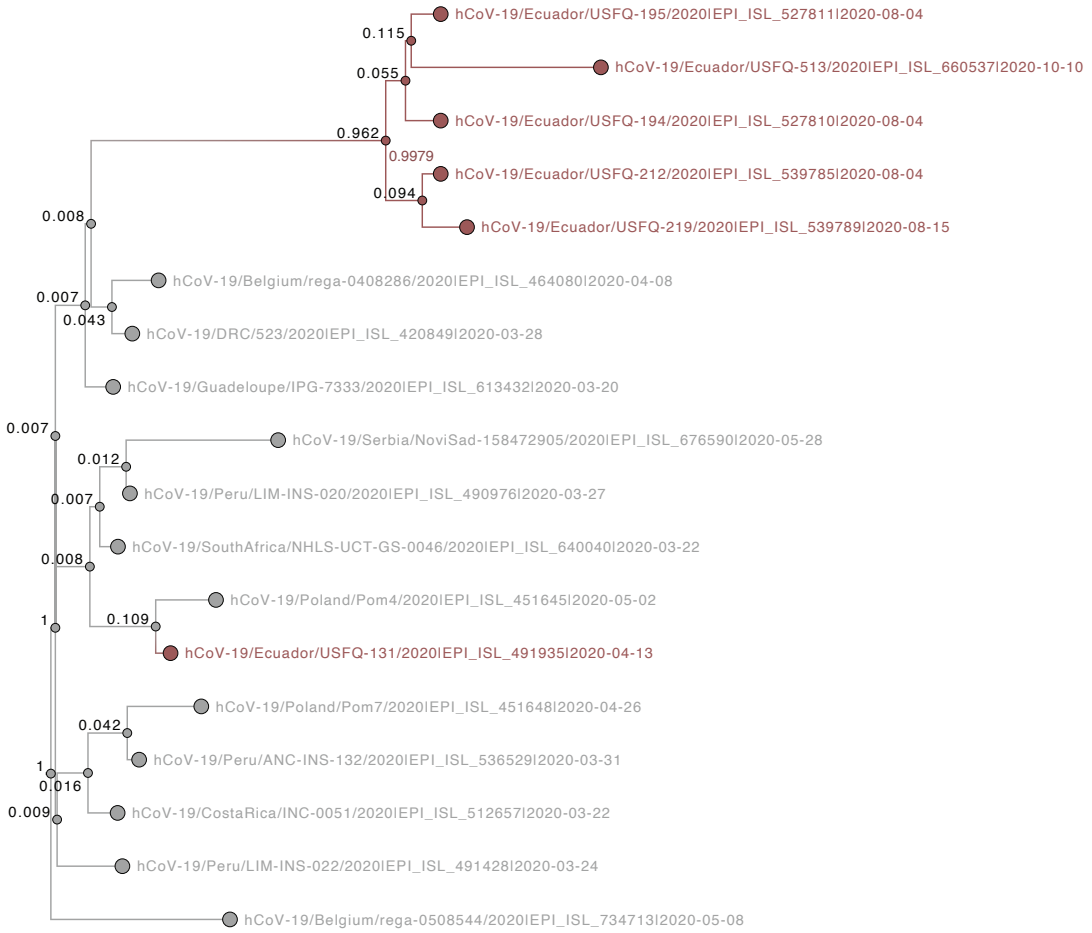


Fig S7. Subtree from the MCC tree, corresponding to transmission lineage F. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

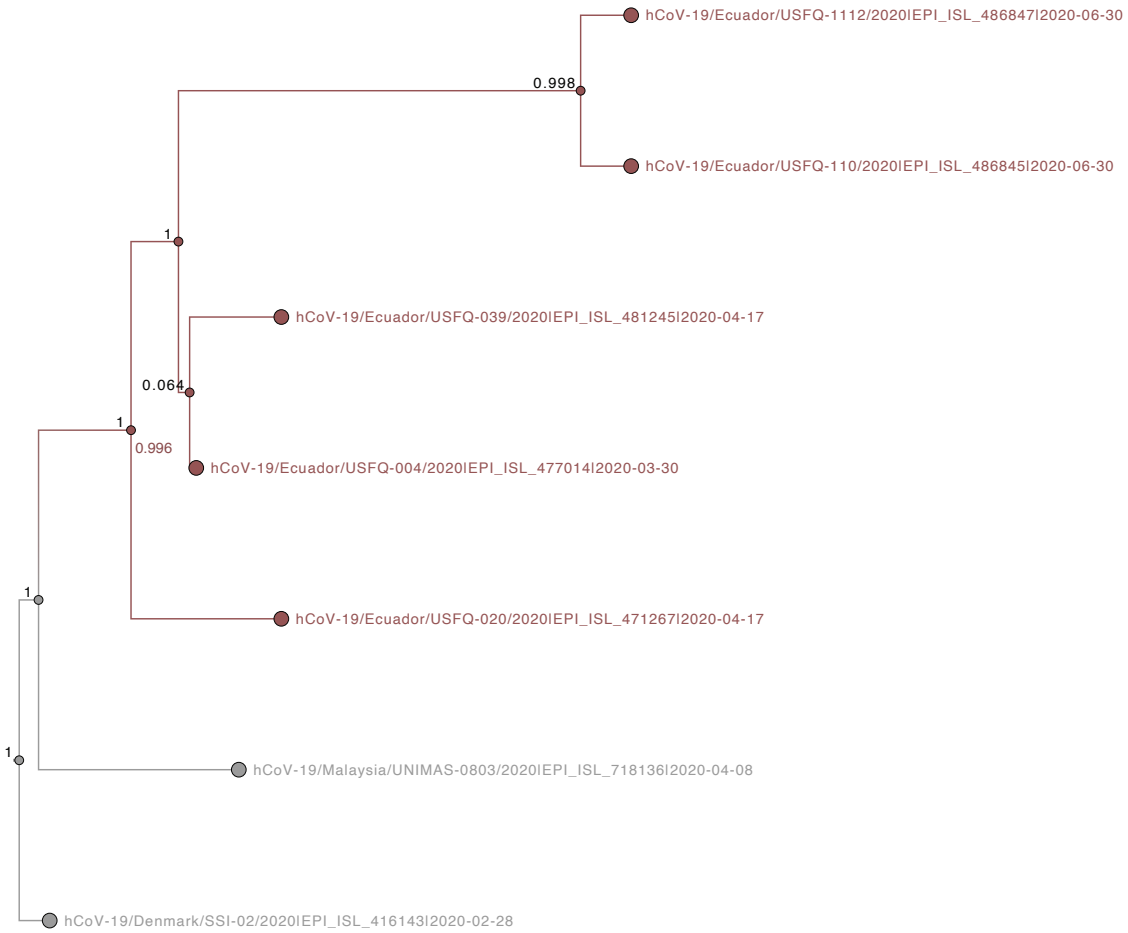


Fig S8. Subtree from the MCC tree, corresponding to transmission lineage B. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

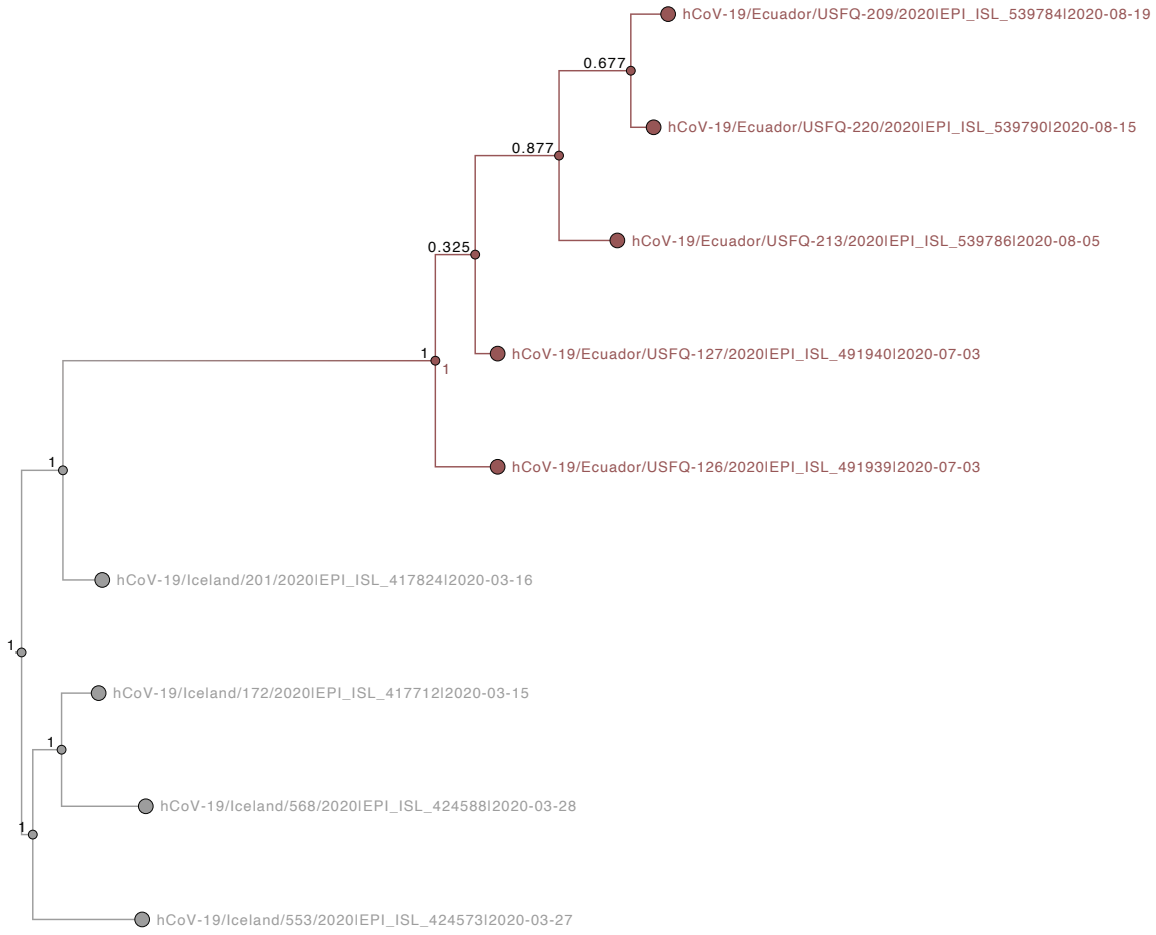


Fig S9. Subtree from the MCC tree, corresponding to transmission lineage M. Posterior node support is shown in black, and location probability support of Ecuador ancestral nodes is shown in red.

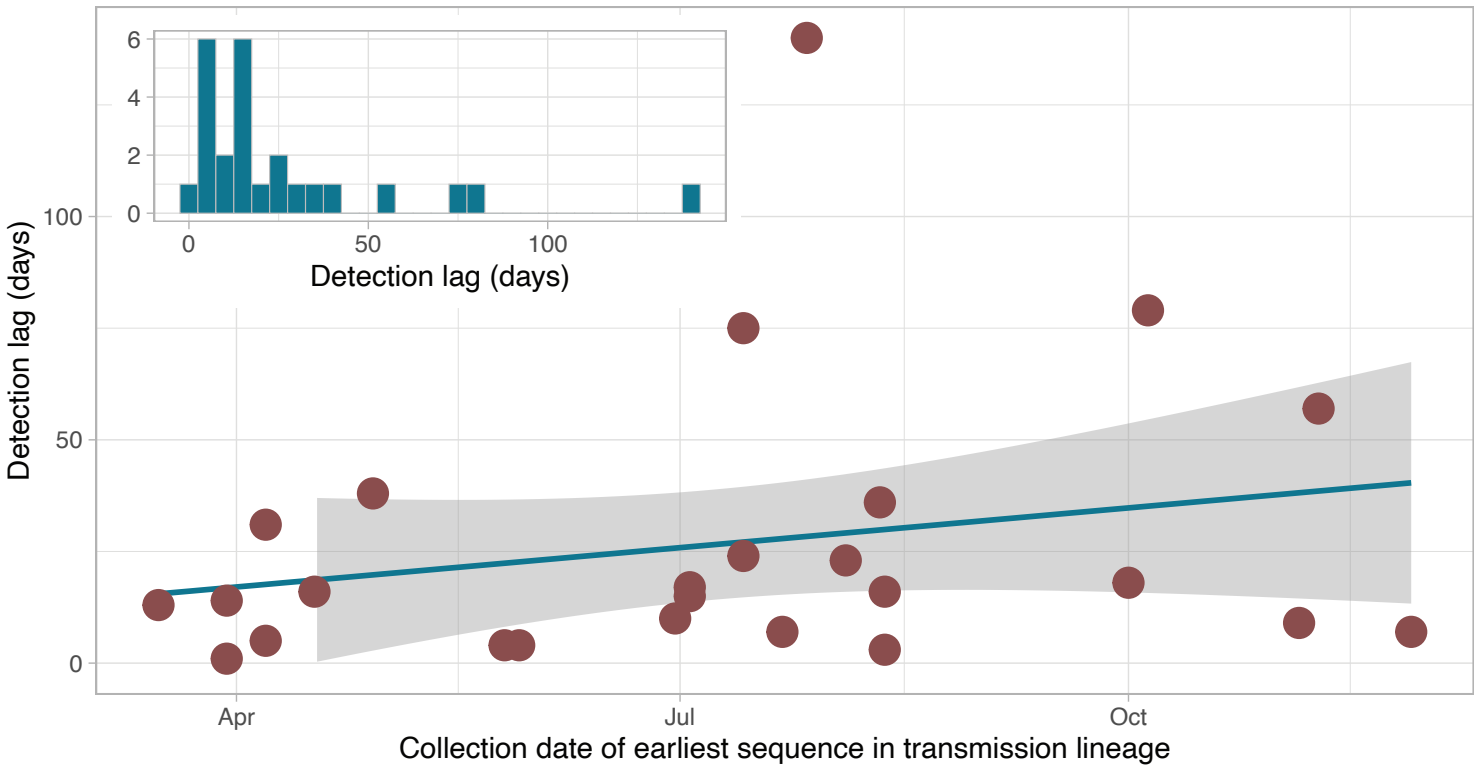


Fig S10. Changes in detection lag over time. Figure shows the detection lag of every transmission lineage (time between the TMRCA of a transmission lineage and the collection date of its earliest sequence), plotted against the collection date of the earliest sequence in said lineage. The inset shows the distribution of detection lag times in the data set.

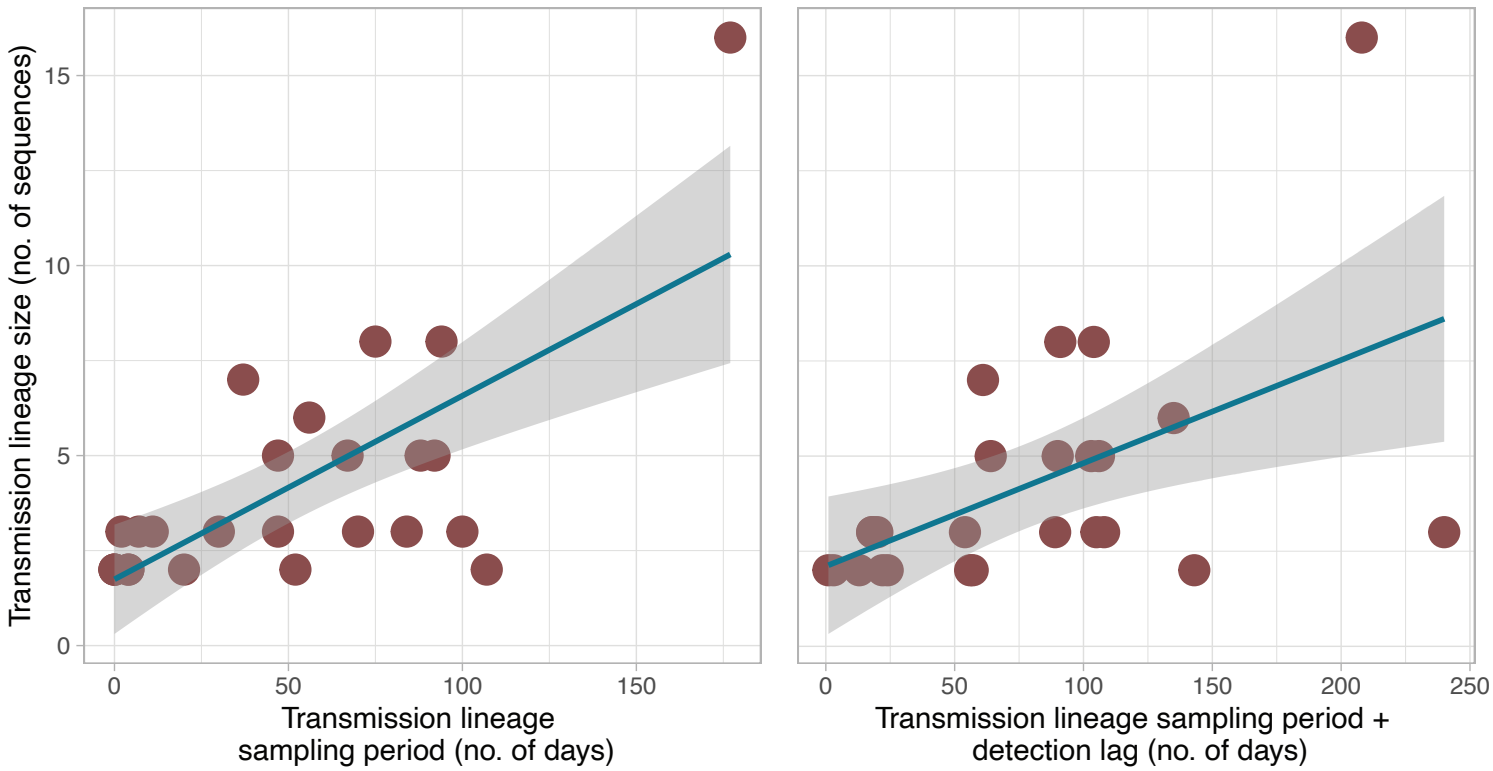


Fig S11. Comparison between the size of transmission lineages (i.e. number of sequences assigned to a transmission lineage) and their persistence in time estimated as either the number of days between the earliest and most recent sequences in said lineage (left) or the number of days between the transmission lineage TMRCA and the most recent sequence (right).

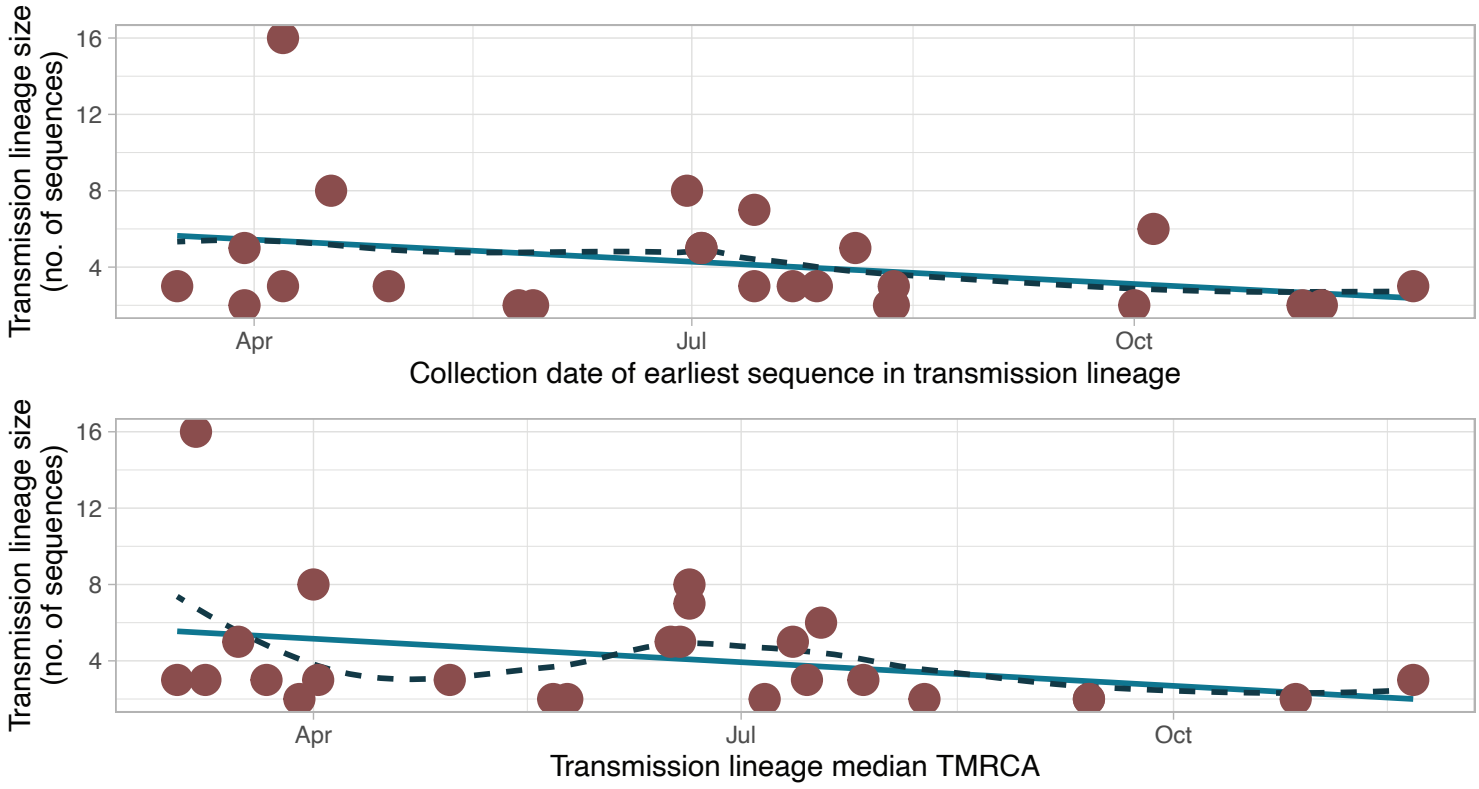


Fig S12. Comparison between the size of transmission lineages (i.e. number of sequences assigned to a transmission lineage) and the collection date of the first sequence identified in said lineage (upper panel) or the TMRCA for said transmission lineage (lower panel). Trend lines show a linear regression (green) and a fitted local polynomial regression (R function stats::loess, black and dashed).

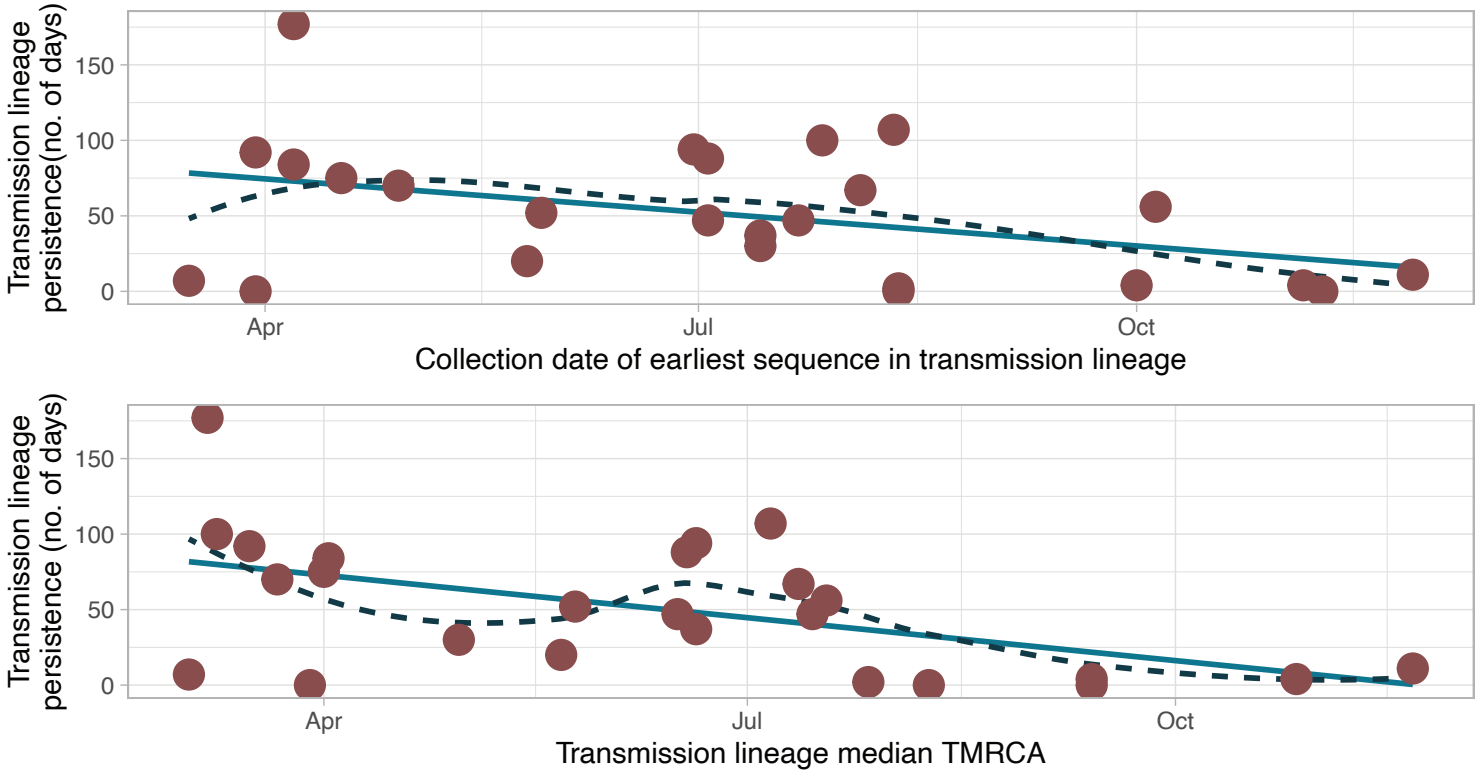


Fig S13. Comparison between the persistence in time of transmission lineages (number of days between the earliest and most recent sequences in said lineage) and the collection date of the first sequence identified in said lineage (upper panel) or the TMRCA for said transmission lineage (lower panel). Trend lines show a linear regression (green) and a fitted local polynomial regression (R function stat::loess, black and dashed).

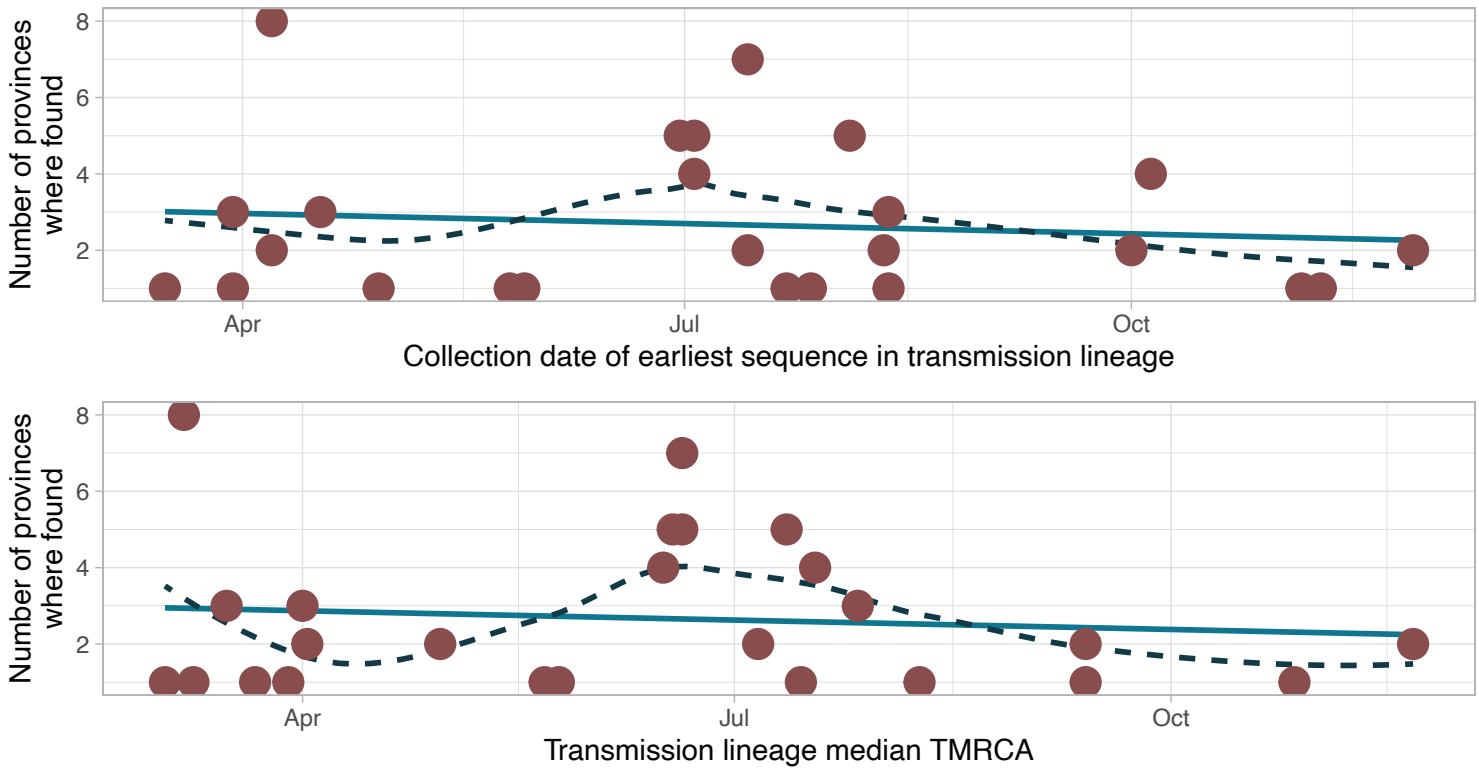


Fig S14. Comparison between the geographic spread of transmission lineages (number of provinces where they've been detected) and the collection date of the first sequence identified in said lineage (upper panel) or the TMRCA for said transmission lineage (lower panel). Trend lines show a linear regression (green) and a fitted local polynomial regression (R function stat::loess, black and dashed).

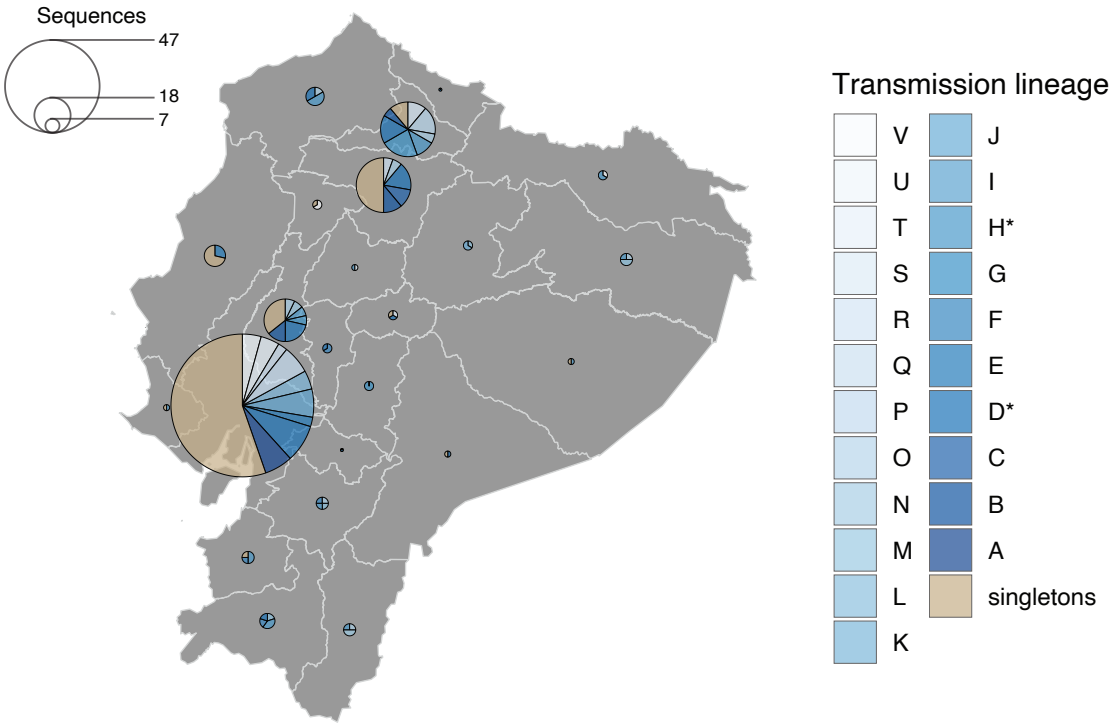


Fig S15. Map of Ecuador showing the contribution of individual transmission lineages (shades of blue) and sequences not associated with them (singletons) in each province. Circle radii represent the number of sequences per province.

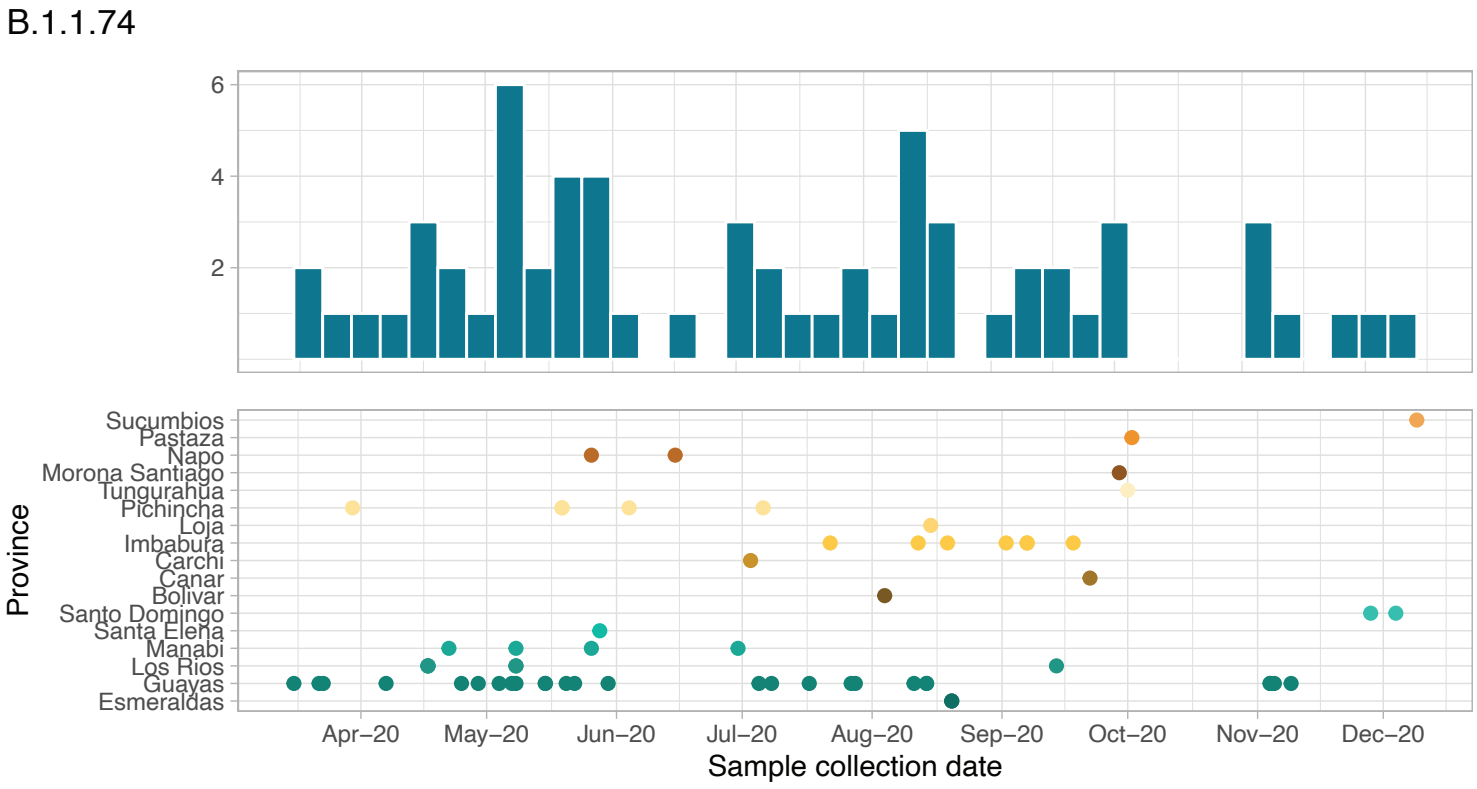


Fig S16. Summary of the identification of Pango lineage B.1.1.74 in Ecuador. The upper panel shows the number of sequences assigned to B.1.1.74 per 7-day time period, and the lower panel shows the province where these sequences were identified by sample collection date. Dot colours correspond to the province where the samples were collected.

Table S1. Summary of sample collection and sequencing methods used by different laboratories in Ecuador.

Laboratory	Sample collection	Whole genome sequencing	Genome assembly
Institute of Virology, Charité-Universitätsmedizin Berlin	Nasopharyngeal swab samples were collected by the SARS-CoV-2 epidemiological surveillance system which were received at the National Reference Center for Influenza and other Respiratory Viruses at INSPI, under cold chain conditions.	Whole-genome amplification was done using the Artic Consortium PCR-based protocol (https://artic.network/ncov-2019). Library preparation and Illumina MiSeq sequencing were done using the KAPA Frag kit and KAPA Hyper Prep kit (Roche Molecular Diagnostics, Switzerland) and MiSeq reagent v2 chemistry (Illumina, USA) according to the manufacturers' protocols.	A modified pipeline dark-matter v3.1.88 (https://pypi.org/project/dark-matter/), using the tools AdapterRemoval, Bowtie2, samtools, gatk4, bcftools, mafft and ivar was used. In addition, the *de novo* assembly produced by Geneious software to the reference SARS-CoV-2 genome Wuhan-Hu-1 (GenBank accession number NC_045512.2) using default, was used to cross-validate as an internal control.
Centre for Multidisciplinary Research of the Direction of Research, Development and Innovation, National Institute of Investigations in Public Health **(INSPI)**	Nasopharyngeal swab samples were collected by the SARS-CoV-2 epidemiological surveillance system which were received at the National Reference Center for Influenza and other Respiratory Viruses at INSPI, under cold chain conditions.	See Lopez-Alvarez *et al.*, 2020.	See Lopez-Alvarez *et al.*, 2020.
Omics Sciences Laboratory, Faculty of Medical Sciences, Universidad de Especialidades Espíritu Santo **(UEES)**	Nasopharyngeal swabs preserved in 400 µL of 1X DNA/RNA Shield™ solution (Zymo Research, USA); saliva, urine, blood and semen samples were used for RNA extraction by Quick RNA viral kit (Zymo Research, CA, USA) according to the manufacturer's protocol in a biosafety type II chamber with HEPA	Nasopharyngeal swab RNA samples were further analyzed with the CleanPlex® SARS-CoV-2 FLEX Panel (Paragon Genomics, Hayward, CA, USA) following the manufacturer's protocol and sequenced on a MiniSeq platform (Illumina, San Diego, CA, USA) with a depth of 1000X.	Assembly and annotation procedures were performed in SOPHIA-DDM-v4

	filters. The RNA extracted was screened for the presence of SARS-CoV-2 genetic material by reverse transcriptase quantitative PCR (RT-qPCR) using Allplex™ SARS-CoV-2 Assay, a multiplex real-time PCR assay to detect 4 target genes: E, N, RdRp and S (Seegene Inc, CA, USA).		
Instituto de Microbiología, Universidad San Francisco de Quito (**IM-USFQ**)	Bronchi-alveolar lavage (BAL), nasopharyngeal swabs; see Márquez *et al.*, 2020.	See Márquez *et al.*, 2020.	See Márquez *et al.*, 2020.
Biomedical Research Unit, Zurita & Zurita Laboratorios (**ZZL**)	Nasopharyngeal swab samples from patients with suspected re-infections.	Whole genome sequencing of the SARS-CoV-2 samples was performed using the Paragon Genomics CleanPlex® SARS-CoV-2 Panel following the manufacturer's instructions on a MiSeq platform (Illumina, San Diego, CA, USA).	FastQC was used to assess the raw reads' quality (Sah *et al.*, 2020) before cleaning them with Trimmomatic 0.39 (PE -phred33 ILLUMINACLIP: TruSeq3-PE.fa:2:30:10:2:keepBothReads (Tillett *et al.*, 2020), LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:40) (Sah *et al.*, 2020). Clean reads were mapped to the reference SARS-CoV-2 genome Wuhan-Hu-1 (GenBank accession number NC_045512.2) (Sah *et al.*, 2020; Tillett *et al.*, 2020), using BWA-MEM 0.7.17 with default parameters (Sah *et al.*, 2020). PCR duplicates were marked and removed with Picard 2.23.8 (Tillett *et al.*, 2020). Variant calling was performed with bcftools 1.11 (-q 25 -Q 35) (Li *et al.*, 2020). Only high-quality variants (QUAL ≥ 20 and DP≥ 5) (Karamitros *et al.*, 2020; Tillett *et al.*, 2020), were retained and

			used to generate consensus genomic sequences. In addition, the *de novo* assembly produced by MEGAHIT 1.2.9 using default parameters (Sah *et al.*, 2020), was used to cross-validate with the reference-based method as an internal control using blastn (Johnson *et al.*, 2008).

Supplementary references

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*((Web Server issue)), W5–W9. https://doi.org/10.1093/nar/gkn201

Karamitros, T., Papadopoulou, G., Bousali, M., Mexias, A., Tsiodras, S., & Mentis, A. (2020). SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *Journal of Clinical Virology*, *131*, 104585. https://doi.org/10.1016/j.jcv.2020.104585

Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., & Jiang, W. (2020). SNPs or RNA modifications? Concerns on mutation-based evolutionary studies of SARSCoV-2. *PLoS ONE*, *15*(8), e0238490. https://doi.org/10.1371/journal.pone.0238490

Lopez-Alvarez, D., Parra, B., Cuellar, W.J. (2020). Genome Sequence of SARS-CoV-2 Isolate Cali-01, from Colombia, Obtained Using Oxford Nanopore MinION Sequencing. *Microbiol Resour Announc*, 9(26), e00573-20. https://doi.org/10.1128/mra.00573-20

Marquez, S., Prado-Vivar, B., Guadalupe, J.J., Gutierrez, B., Jibaja, M., Tobar, M., Mora, F., Gaviria, J., García, M., Espinosa, F., Ligña, E., Reys, J., Barragán, V., Rojas-Silva, P., Trueba, G., Grunauer, M., Cárdenas, P. (2020). Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19 in Ecuador. *medRxiv*, https://dx.doi.org/10.1101%2F2020.06.11.20128330

Sah, R., Rodriguez-Morales, A. J., Jha, R., Chu, D. K. W., Gu, H., Peiris, M., Bastola, A., Lal, B. K., Ojha, H. C., Rabaan, A. A., Zambrano, L. I., Costello, A., Morita, K., Pandey, B. D., & Poon, L. L. M. (2020). Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Isolated in Nepal. *Microbiology Resource Announcements*, *9*(11), e00169-20. https://doi.org/10.1128/mra.00169-20

Tillett, R. L., Sevinsky, J. R., Hartley, P. D., Kerwin, H., Crawford, N., Gorzalski, A., Laverdure, C., Verma, S. C., Rossetto, C. C., Jackson, D., Farrell, M. J., Van Hooser, S., & Pandori, M. (2020). Genomic evidence for reinfection with SARS-CoV-2: A case study. *The Lancet Infectious Diseases*, *21*(1), 52–58. https://doi.org/10.1016/S1473-3099(20)30764-7

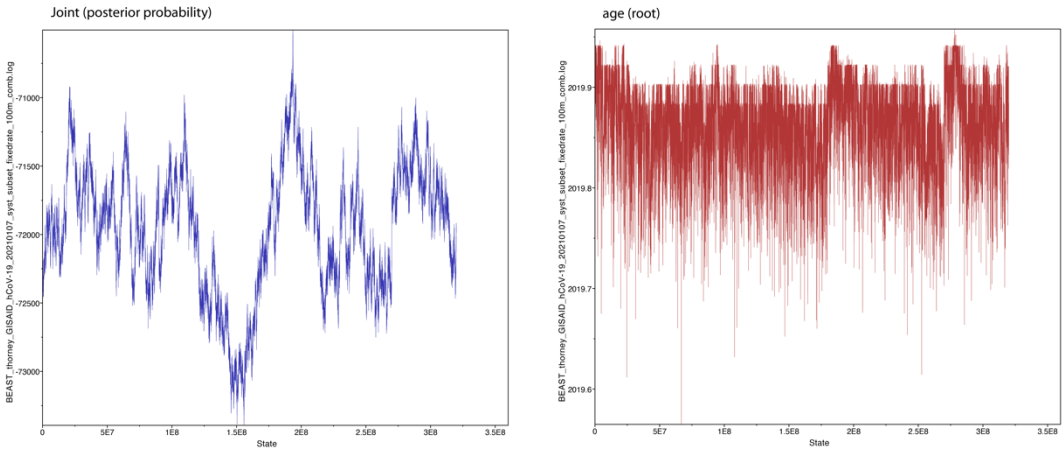


Fig S16. Selected trace plots showing the posterior probability and the tree root age from six combined MCMC chains from the Thorney BEAST software for the systematically downsampled data set.

Appendix 4 Supplementary material for Chapter 4

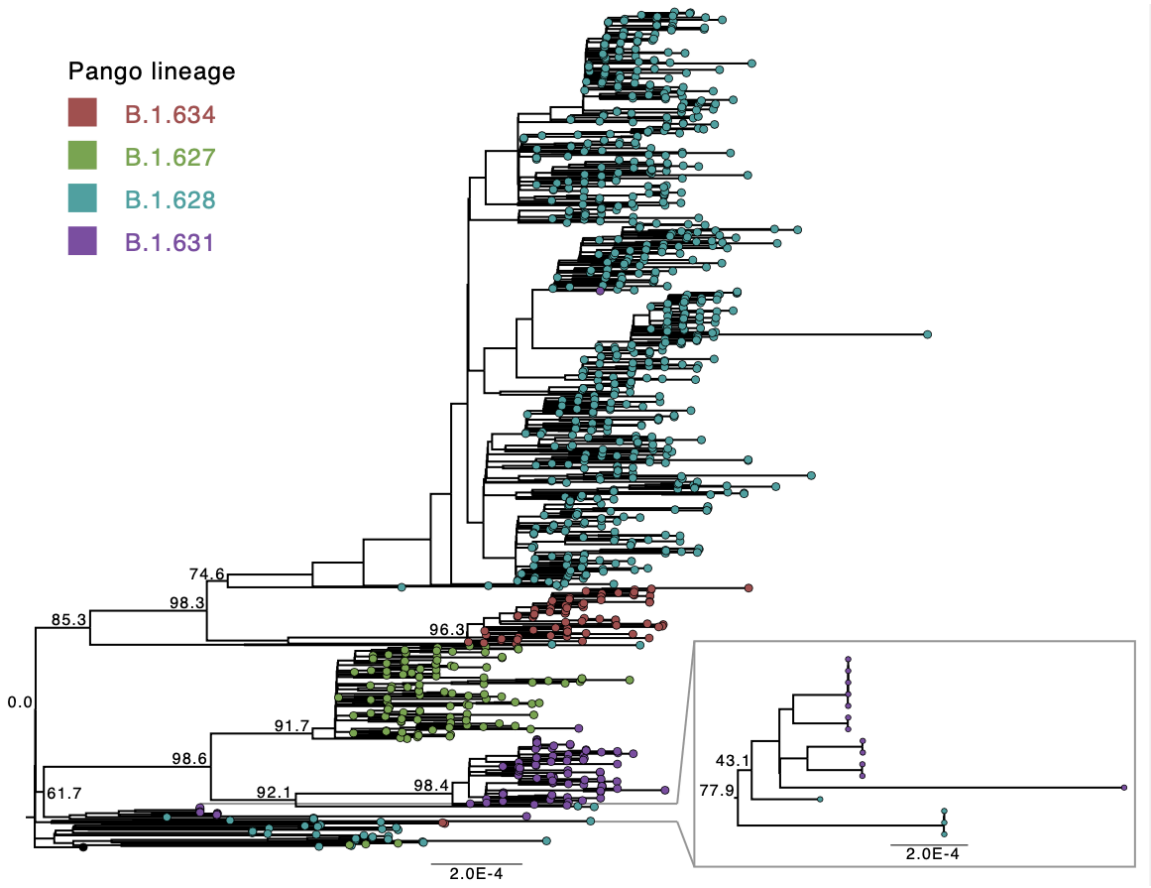


Figure S1. Maximum likelihood phylogenetic analyses of the complete SARS-CoV-2 genome for the four lineages. The individually designated PANGO lineage for each sequence is highlighted (the predominant lineage for sections of the tree shown), and SH-aLRT node support is shown for key lineage defining nodes on both phylogenies.

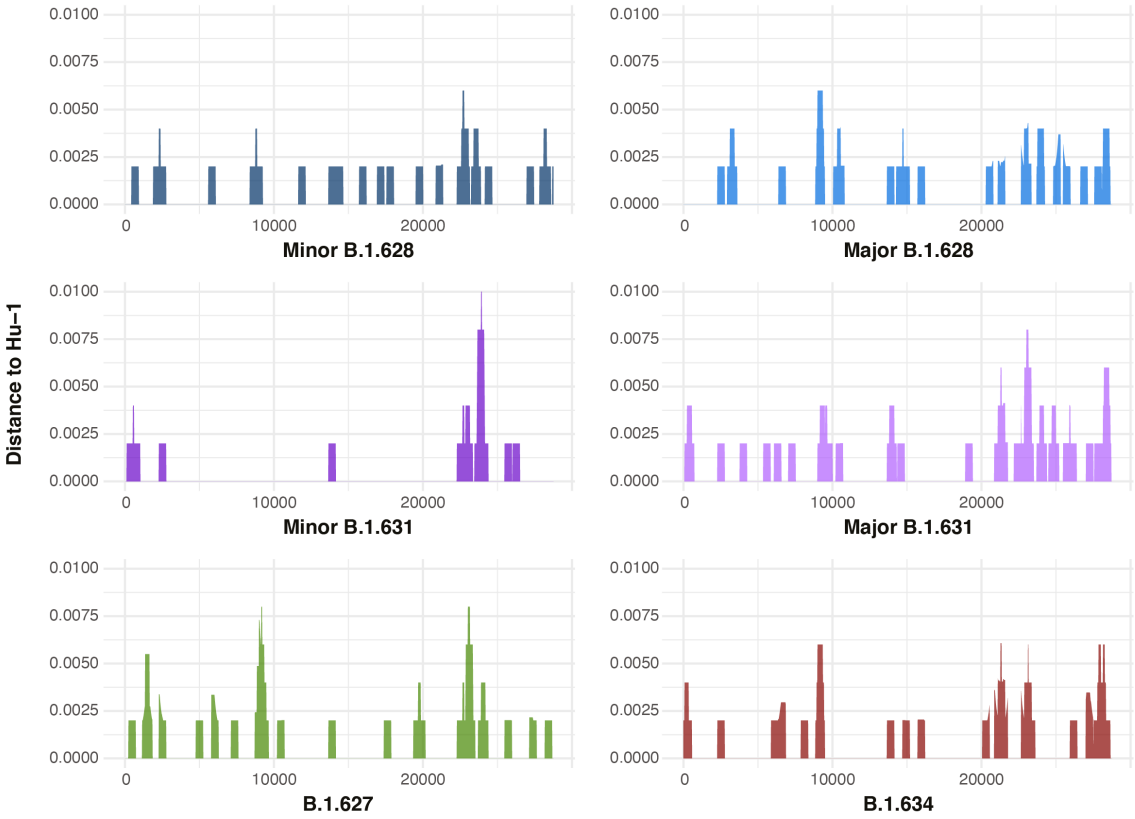


Figure S2. Genetic distance plots across the SARS-CoV-2 genome between the main SARS-CoV-2 phylogenetic clusters amongst the four lineages under investigation and the reference Wuhan Hu-1. Raw genetic distances between basal sequences for each of the monophyletic groups identified in the phylogenetic analyses and the 2019 Wuhan Hu-1 reference genome (MN908947.3). Distances were estimated from genomic segments of 500 nucleotides in length, overlapping over 20-nucleotide intervals.

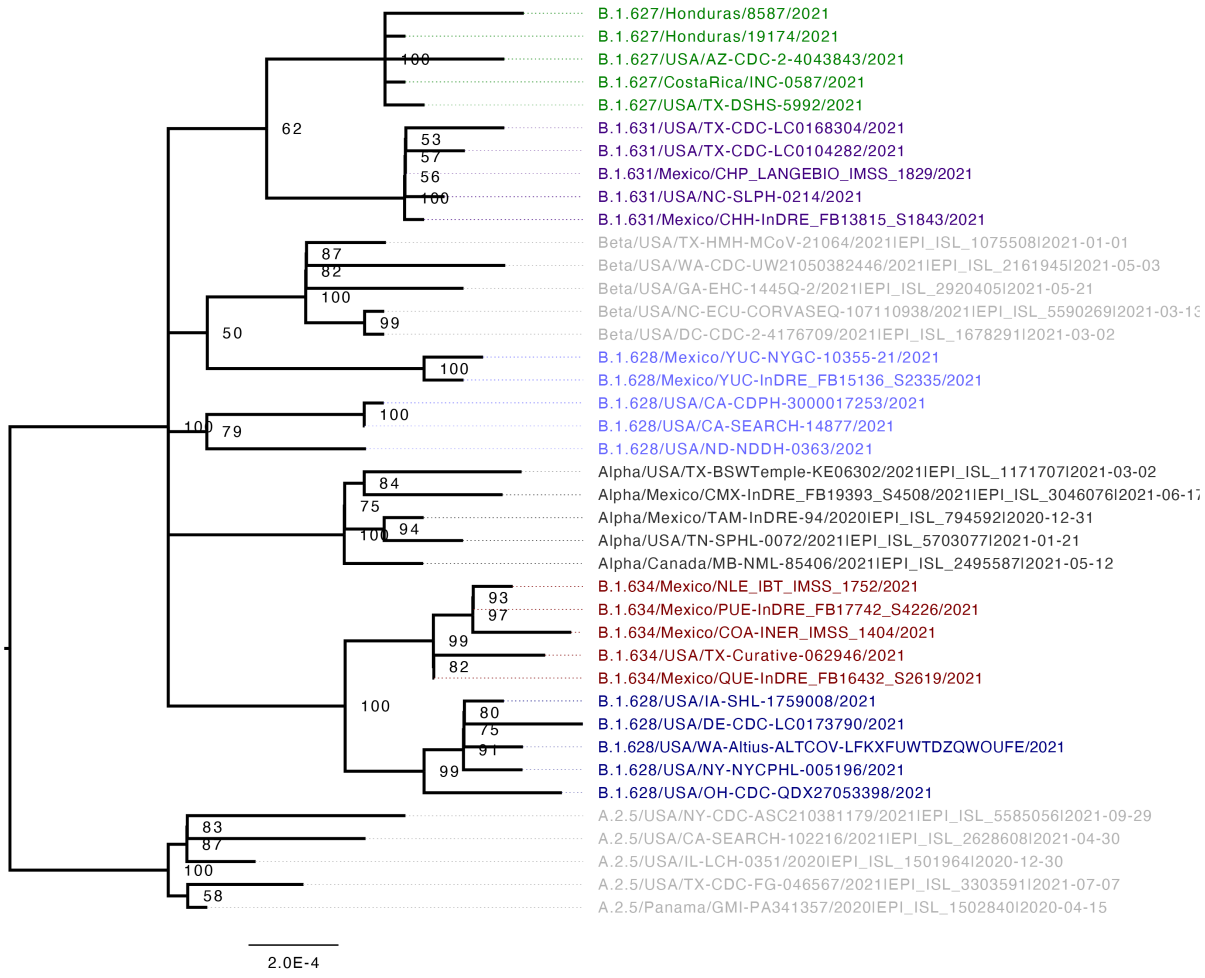


Figure S3. Maximum likelihood phylogenetic tree for the Orf1ab genome segment of a selection of sequences of lineages under investigation B.1.627 (green), B.1.628 minor (light blue), B.1.628 major (dark blue), B.1.631 (purple) and B.1.634 (red) in relation to B.1.1.7 (VOC Alpha, dark grey) and B.1.351 (VOC Beta, fark grey). Node support is shown from 1000 bootstrap replicates, nodes with support <50% are collapsed into polytomies. The tree is rooted in reference to lineage A.2.5.

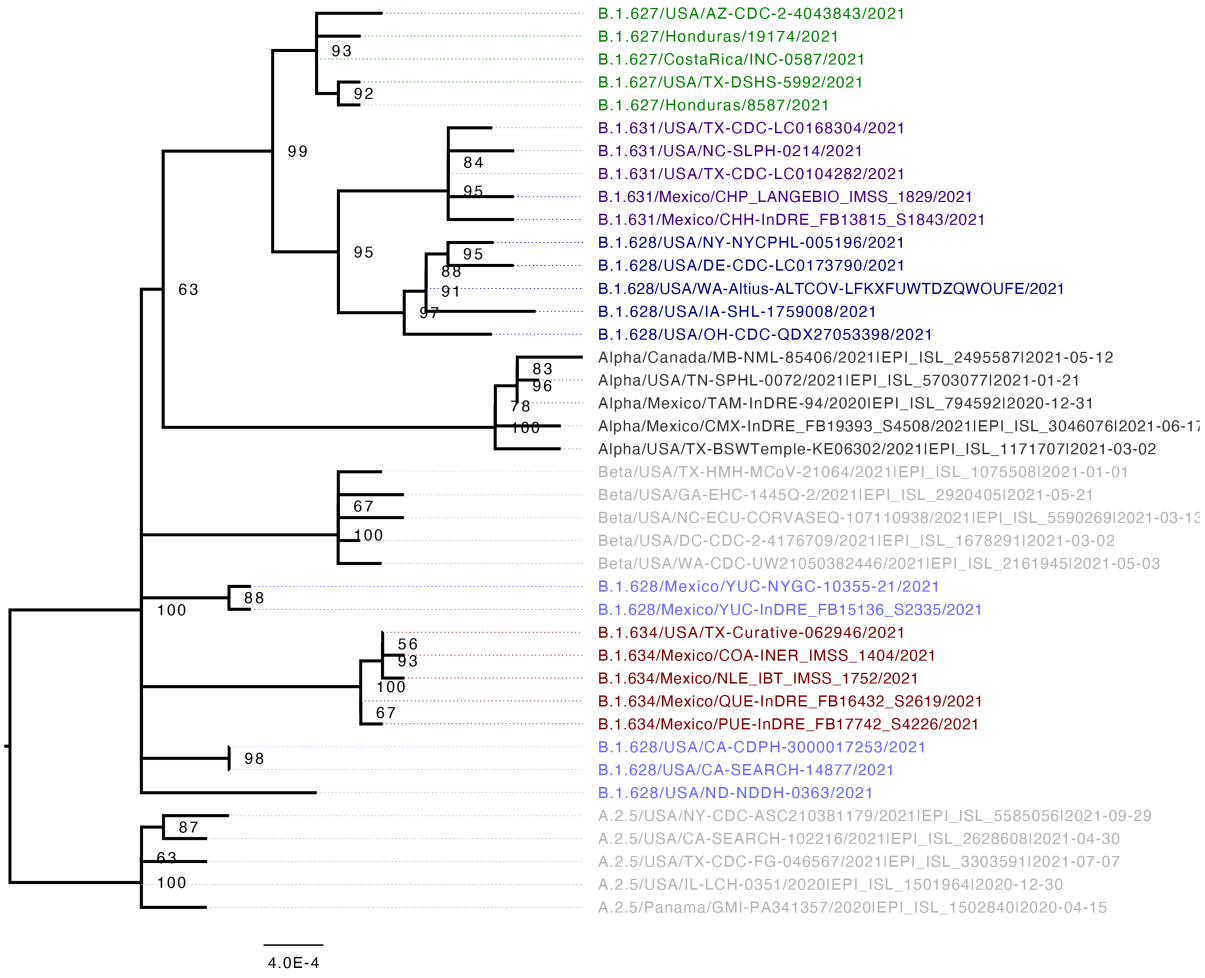


Figure S4. Maximum likelihood phylogenetic tree for the S-3' genome segment of a selection of sequences of lineages under investigation B.1.627 (green), B.1.628 minor (light blue), B.1.628 major (dark blue), B.1.631 (purple) and B.1.634 (red) in relation to B.1.1.7 (VOC Alpha, dark grey) and B.1.351 (VOC Beta, fark grey). Node support is shown from 1000 bootstrap replicates, nodes with support <50% are collapsed into polytomies. The tree is rooted in reference to lineage A.2.5.

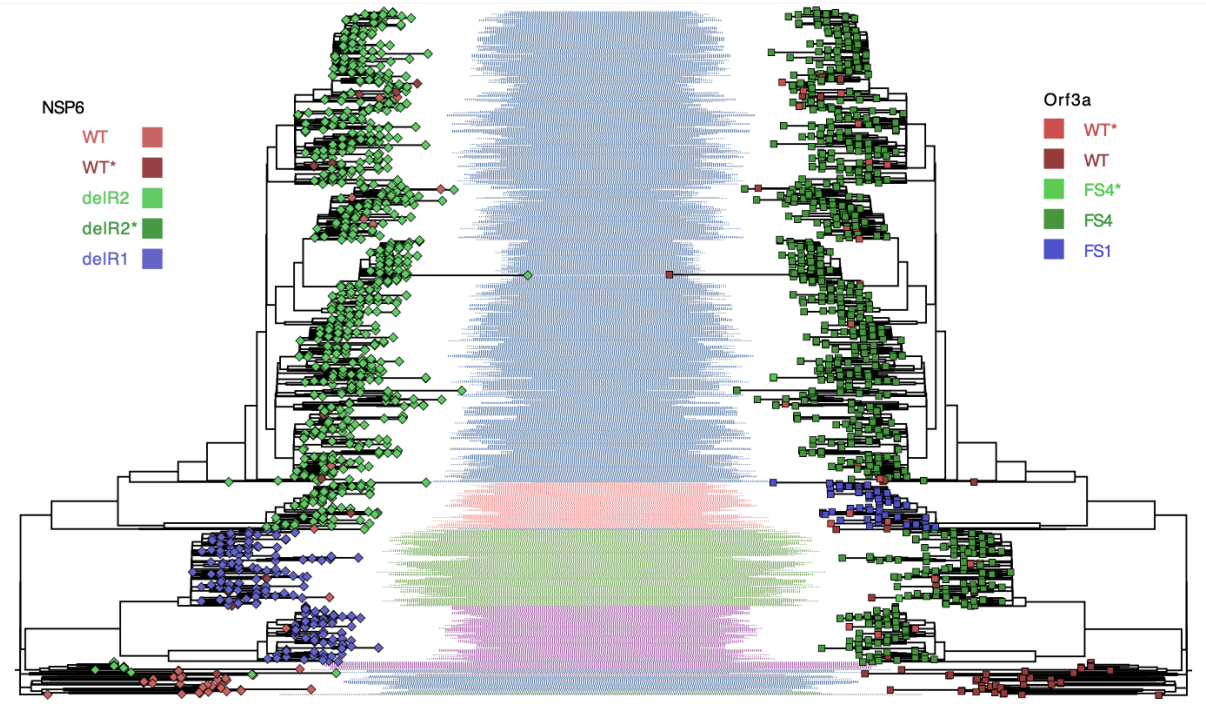


Figure S5. Deletions on the Orf1ab (NSP6; diamonds) and Orf3a (rectangles) loci mapped to the maximum likelihood phylogeny of the complete genome for the four lineages under investigation. Major lineage designations are shown in coloured shading: B.1.627 (green), B.1.628 (blue), B.1.631 (purple) and B.1.634 (red). Darker colours assigned to tips correspond to sequences where the marked deletion occurs amongst ambiguous sites (Ns).

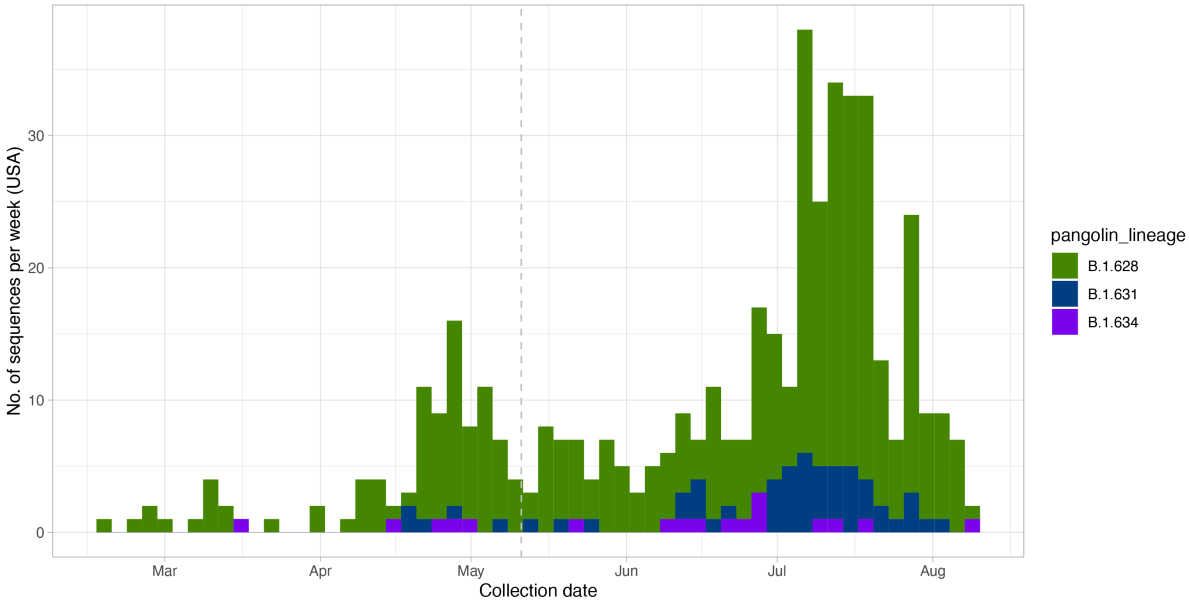


Figure S6. Sequence sampling from the B.1.628 major, B.1.631 and B.1.634 lineages in the USA. Sequences shown here correspond only to the sequences included in the phylogenetic analyses (i.e. <10% ambiguities in the genome sequence, >90% genome coverage). Dotted line shows the starting date for the systematic genomic surveillance work performed in Mexico by the CoVi-Gen Mex Consortium.

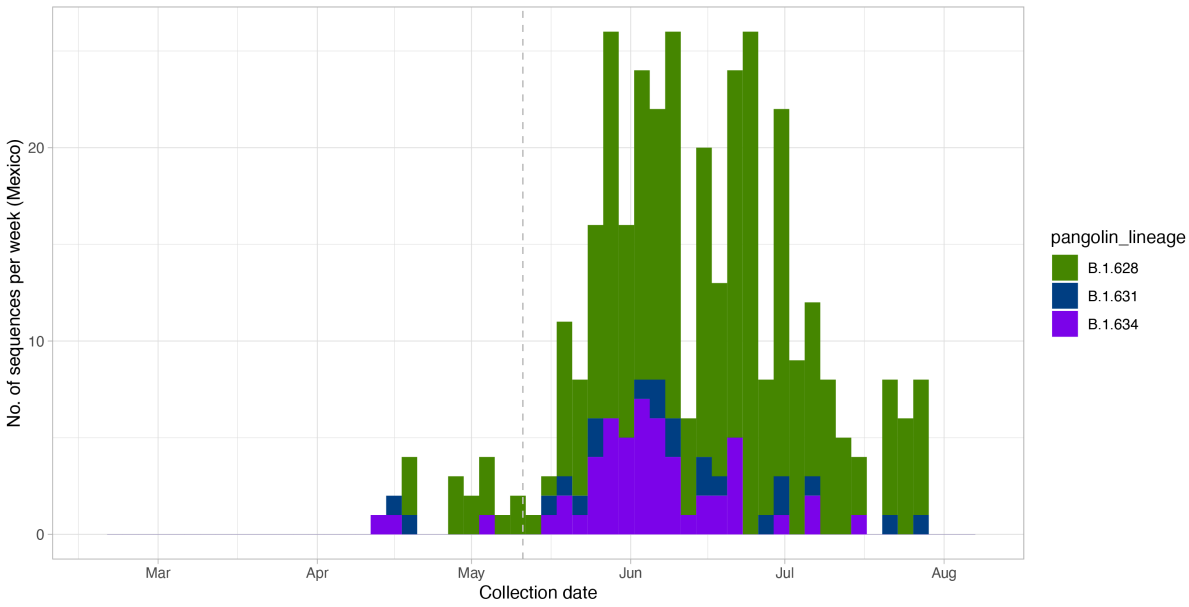


Figure S7. Sequence sampling from the B.1.628 major, B.1.631 and B.1.634 lineages in Mexico. Sequences shown here correspond only to the sequences included in the phylogenetic analyses (i.e. <10% ambiguities in the genome sequence, >90% genome coverage). Dotted line shows the starting date for the systematic genomic surveillance work performed in Mexico by the CoVi-Gen Mex Consortium.

Appendix 5 Supplementary material for Chapter 5

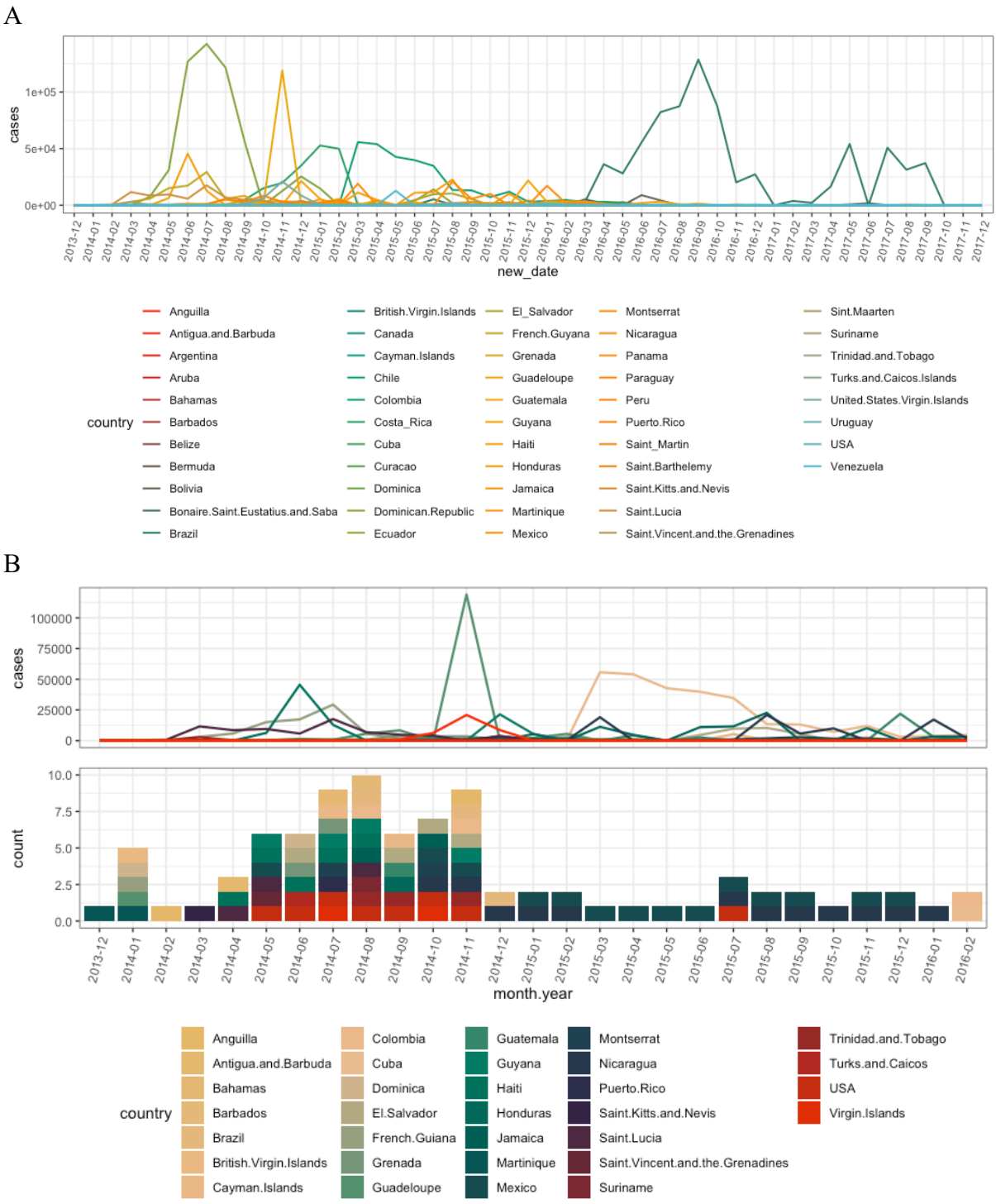


Figure S1. CHIKV epidemiological and genomic surveillance trends in the Americas. ***(A)*** Monthly number of CHIKV cases reported to the Pan-American Health Organisation (PAHO) between 2013 and 2017, grouped by country. ***(B)*** Comparison between monthly number of cases (reported to PAHO, upper panel) in countries that have generated CHIKV genome sequences and publicly available complete CHIKV

genome sequences (lower panel). Mexico sequences include those generated in this study.

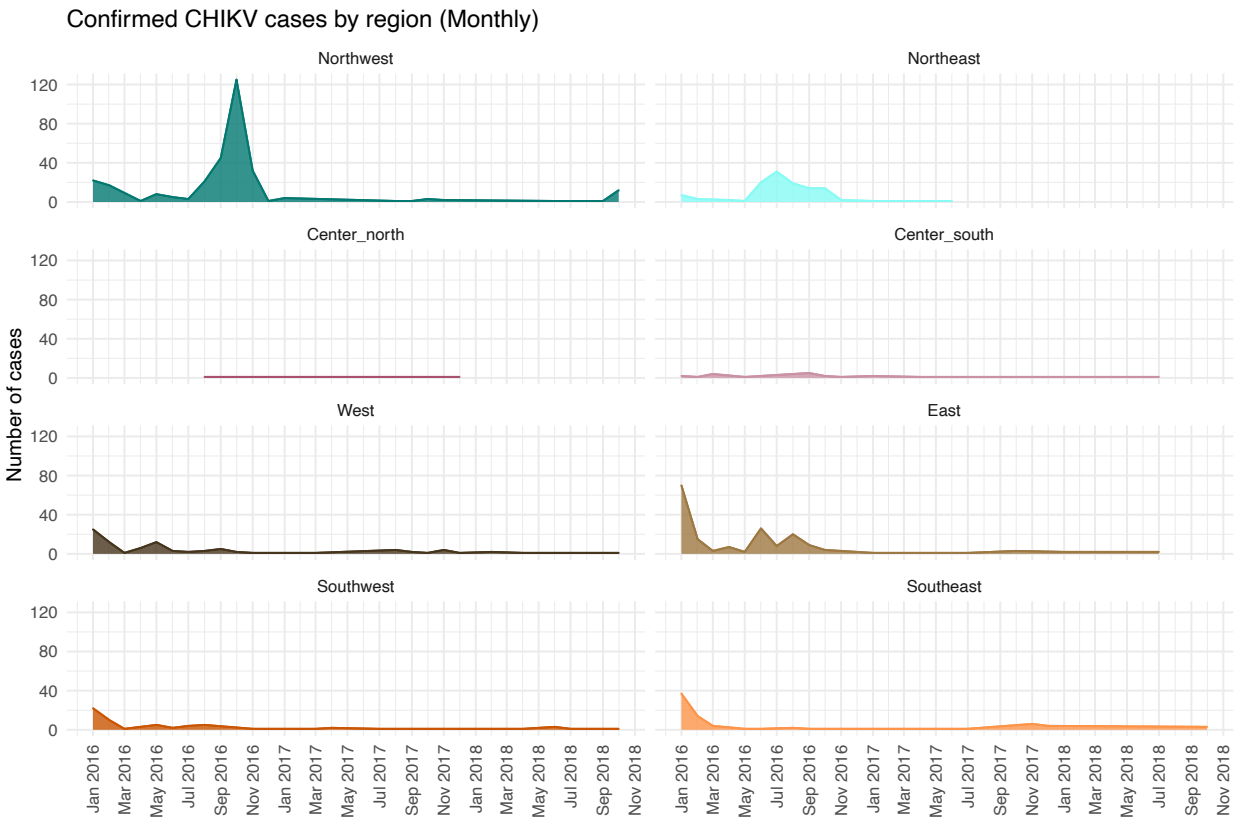


Figure S2. CHIKV epidemiological trends across Mexico between 2016 and 2018. Monthly cases reported to the SINAVE surveillance system (InDRE/Ministry of Health).

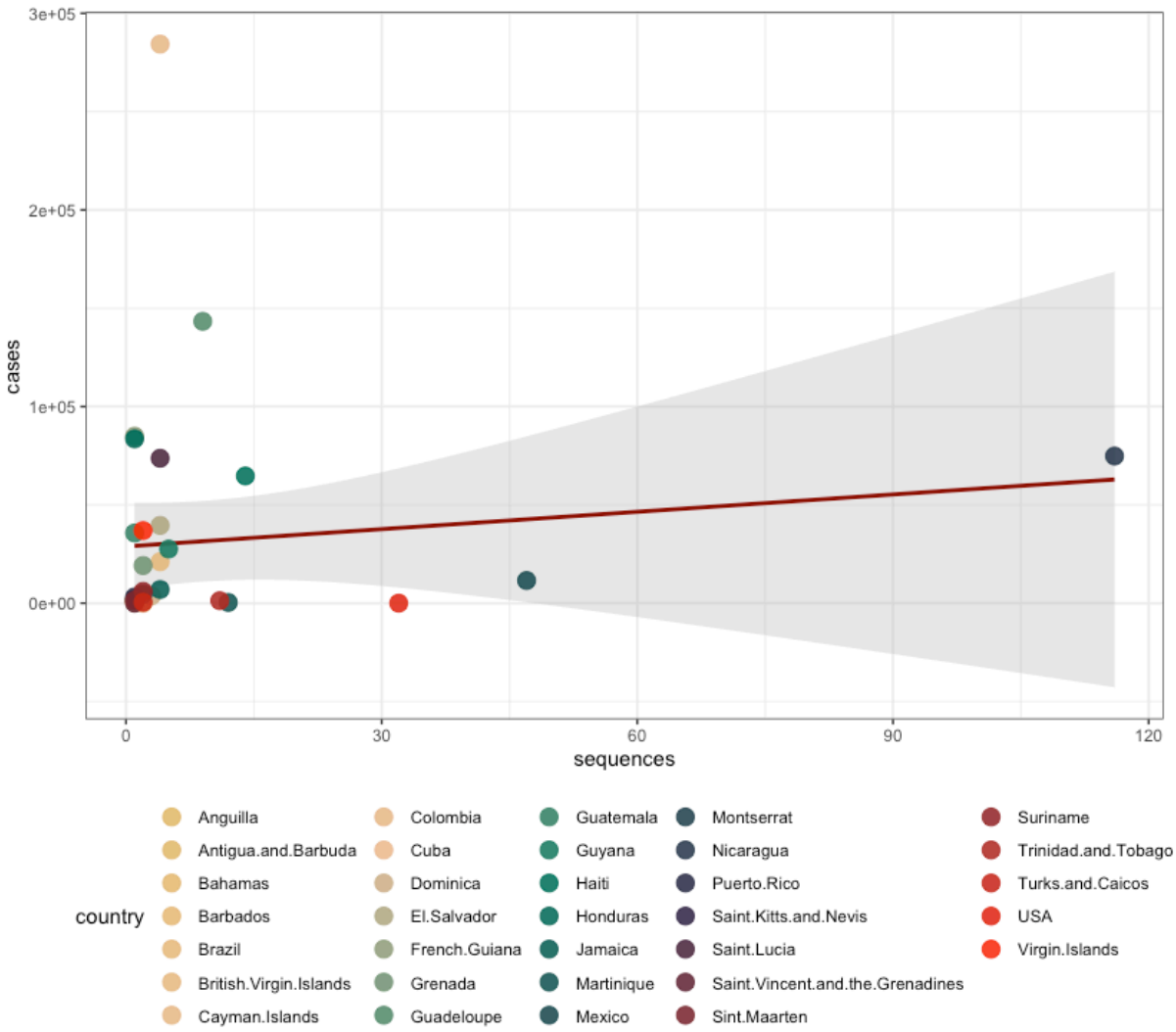


Figure S3. Complete CHIKV genome sequences versus total cases reported to PAHO per country. Regression 95% confidence interval shown in grey.

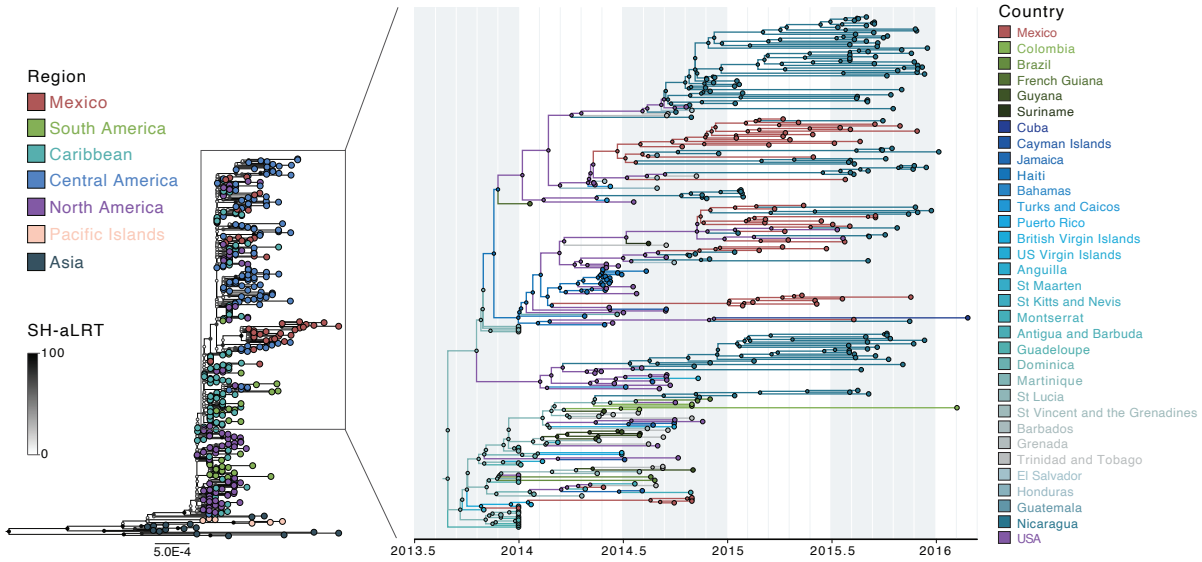


Figure S4. Phylogenetic analyses of CHIKV in the Americas. Left panel shows a ML phylogenetic tree of all CHIKV complete genome sequences from the Americas included in our analysis. Tree tips are coloured by region where they were collected, and node shapes are coloured by their statistical support. Right panel shows a time-calibrated MCC tree for the well-defined CCNA clade. Tips and branches are coloured by their inferred country of collection and circulation respectively, as inferred through a DTA phylogeographic analysis.

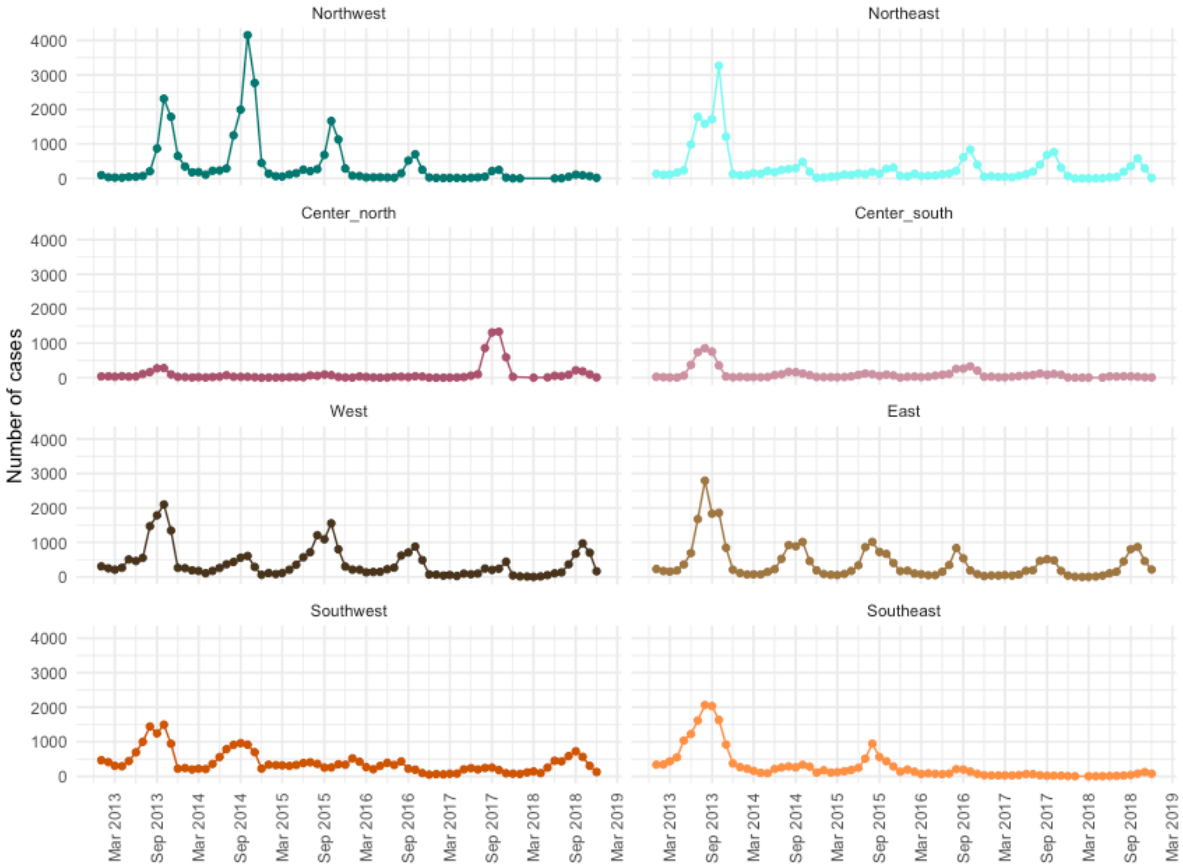


Figure S5. DENV epidemiological trends across Mexico between 2016 and 2018. Monthly cases reported to the SINAVE surveillance system (InDRE/Ministry of Health), aggregating dengue fever (DF) and dengue haemorrhagic fever (DHF).

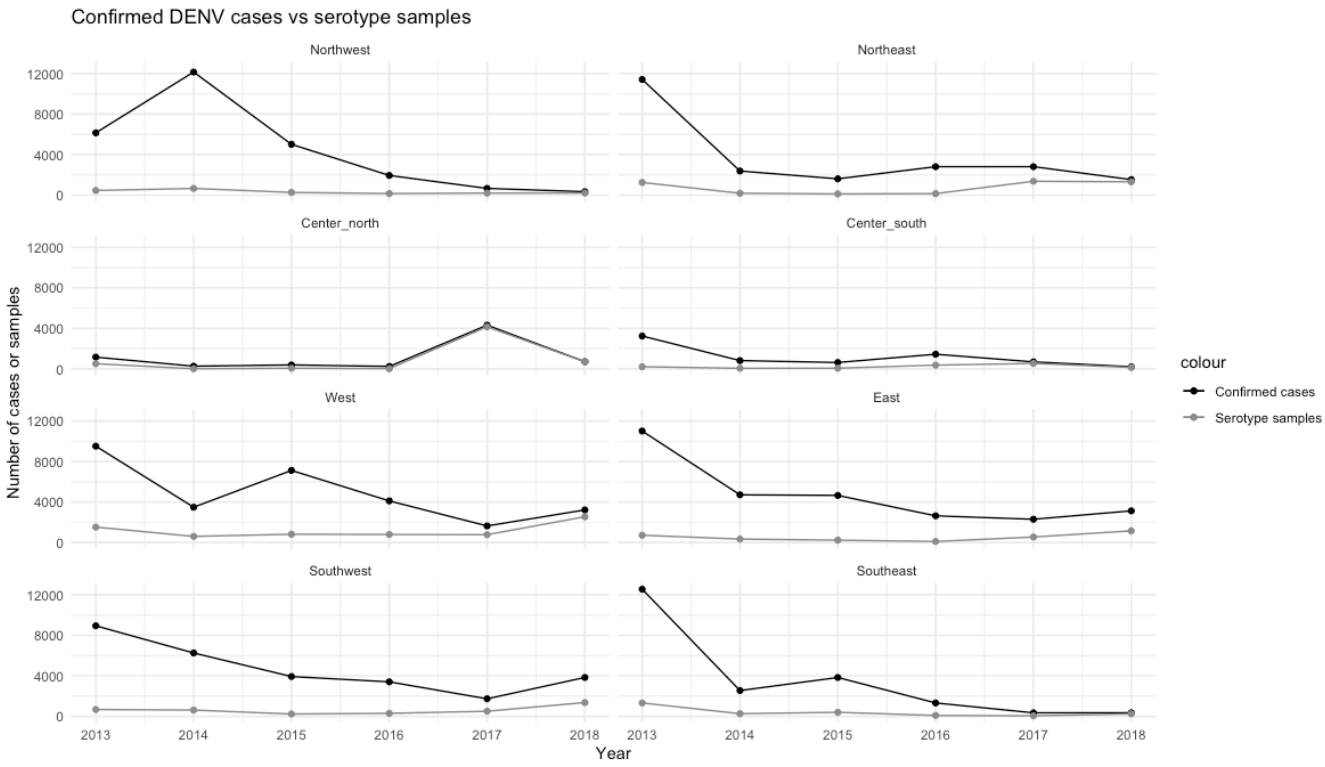


Figure S6. Serotyping representation for DENV across regions in Mexico. Total number of reported cases by year (black line) compared to total number of cases where the causal serotype has been identified and reported (grey line), between 2013 and 2018. Data from the SINAVE surveillance system (InDRE/Ministry of Health).

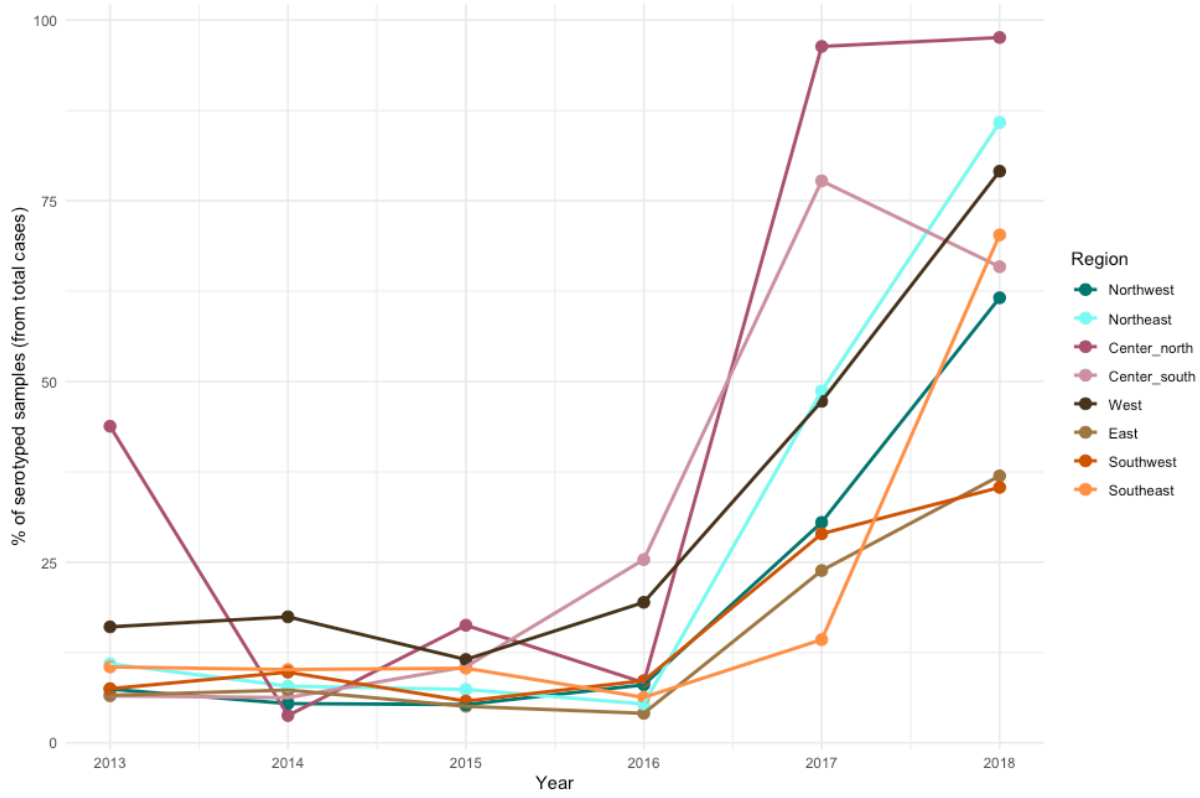


Figure S7. DENV serotyping efficacy across time in different Mexico regions. Percentage of serotyped samples per region per year between 2013 and 2018.

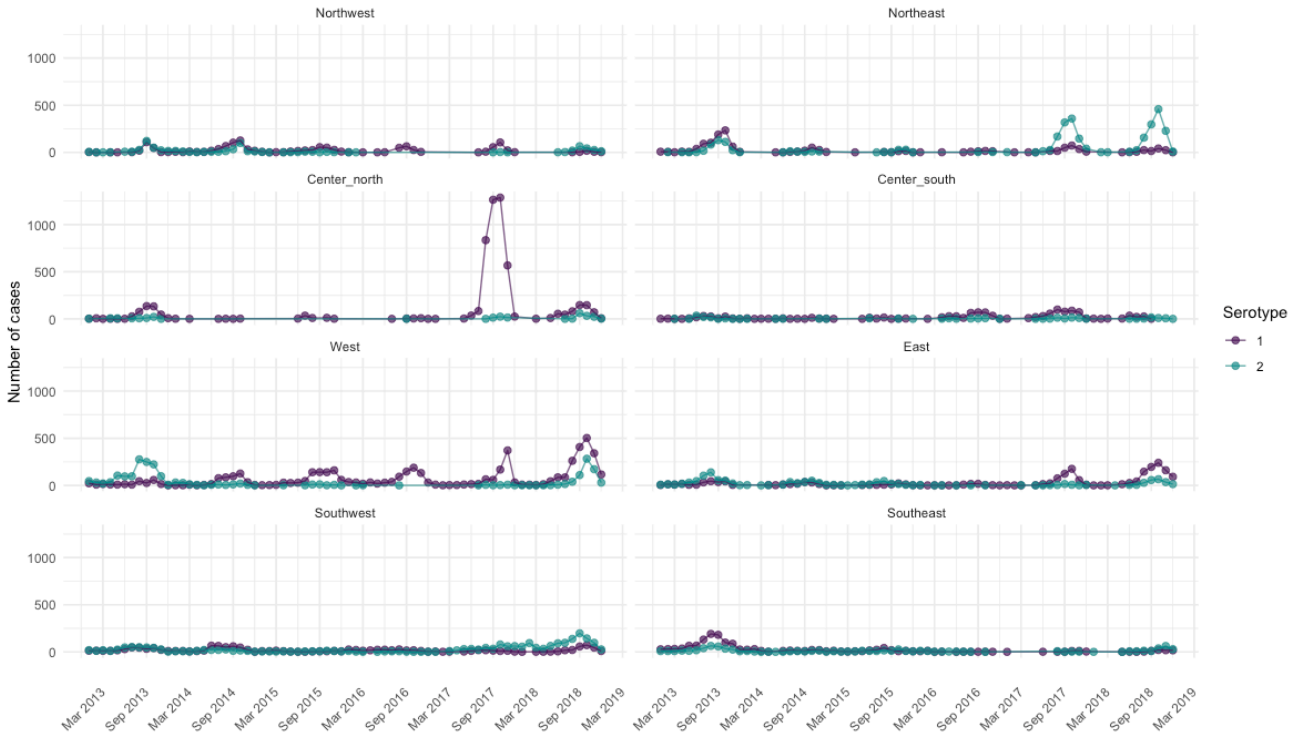


Figure S8. DENV-1 and DENV-2 case numbers across Mexico regions. Monthly numbers of cases identified as DENV-1 (purple) or DENV-2 (teal) between 2013 and 2018 across geographic regions in the country.

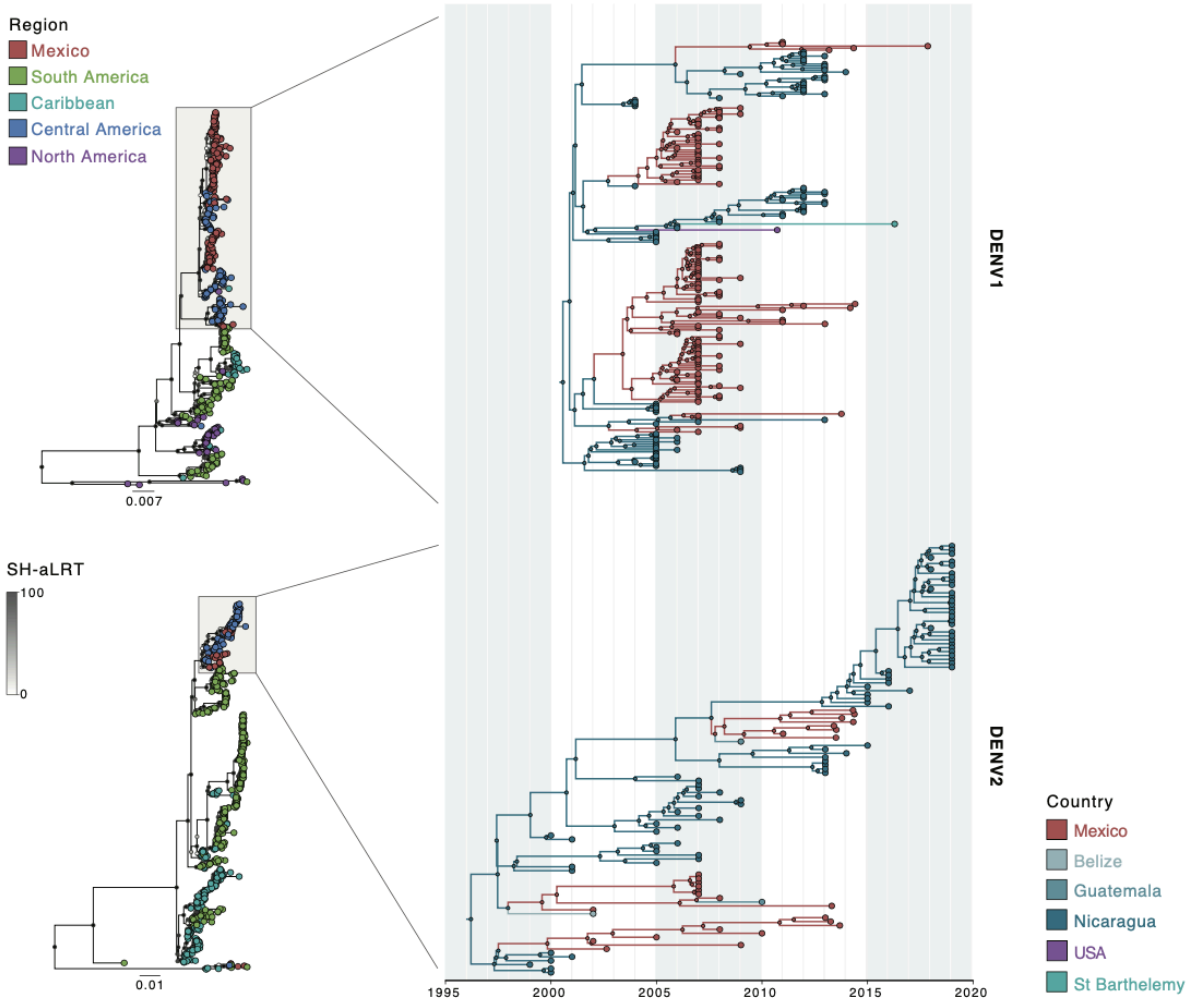


Figure S9. Phylogenetic analyses of DENV-1 and DENV-2 in the Americas. Left panel shows ML phylogenetic trees of all DENV-1 (upper) and DENV-2 (lower) complete genome sequences from the Americas included in our analysis. Tree tips are coloured by region where they were collected, and node shapes are coloured by their statistical support. Right panel shows time-calibrated MCC trees for the well-defined CCNA clades for each virus. Tips and branches are coloured by their inferred country of collection and circulation respectively, as inferred through a DTA phylogeographic analysis.

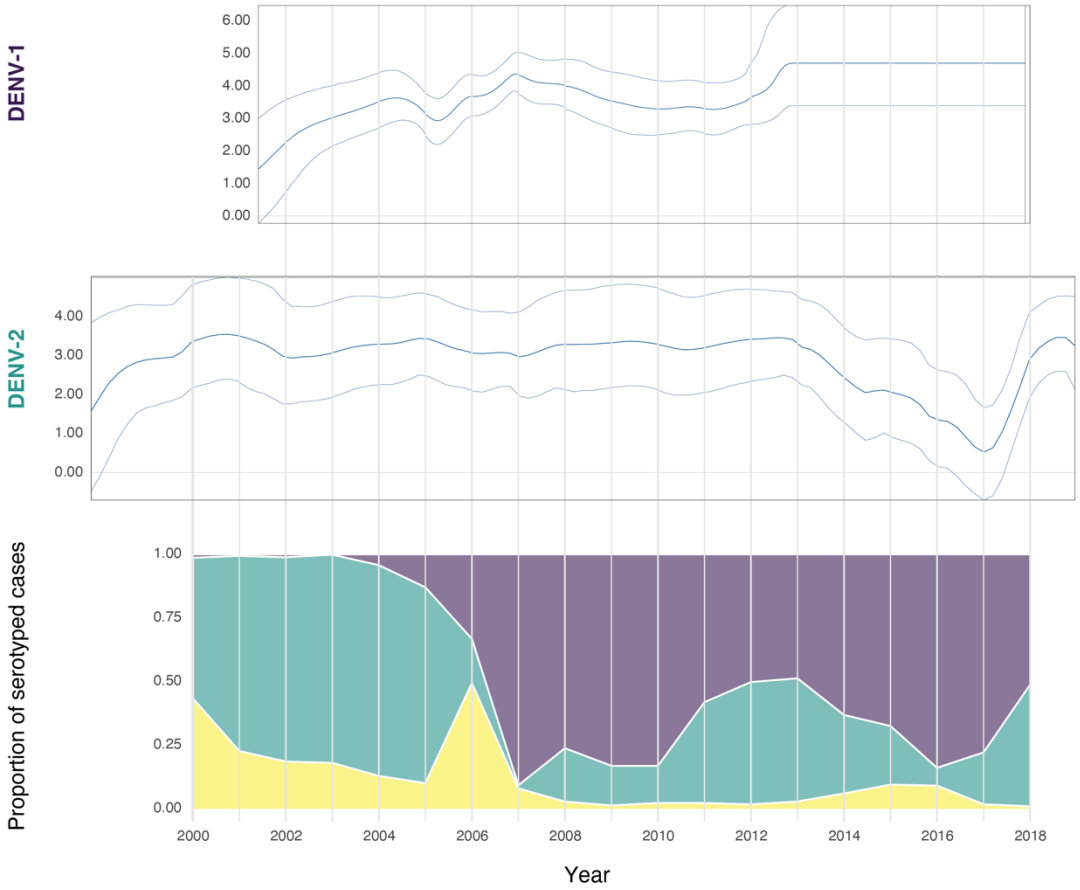


Figure S10. Bayesian Skyline plots of DENV-1 and DENV-2 effective population sizes over time. Plots were generated from a GMRF Bayesian Skyride model and co-estimated with DENV molecular phylogenies of the CCNA clades in BEAST. A plot of the serotype frequencies (obtained from the SINAVE) are shown below for reference.

Table S1. PAHO regions

PAHO region	Country
North America	USA
North America	Mexico
North America	Canada
North America	Bermuda
Central America	Panama
Central America	Nicaragua
Central America	Honduras
Central America	Guatemala
Central America	El Salvador
Central America	Costa Rica
Central America	Belize
Latin America	Saint Martin
Latin America	Martinique
Latin America	Guadeloupe
Latin America	Puerto Rico
Latin America	Haiti
Latin America	Guyana
Latin America	French Guyana
Latin America	Dominican Republic
Latin America	Cuba
Andean	Venezuela
Andean	Peru
Andean	Ecuador
Andean	Colombia
Andean	Bolivia
South Cone	Uruguay
South Cone	Paraguay
South Cone	Chile
South Cone	Brazil
South Cone	Argentina
Non-Latin Caribbean	Saint Barthelemy
Non-Latin Caribbean	Sint Maarten
Non-Latin Caribbean	United States Virgin Islands
Non-Latin Caribbean	Turks and Caicos Islands
Non-Latin Caribbean	Trinidad and Tobago
Non-Latin Caribbean	Suriname
Non-Latin Caribbean	Saint Vincent and the Grenadines
Non-Latin Caribbean	Saint Lucia
Non-Latin Caribbean	Saint Kitts and Nevis
Non-Latin Caribbean	Montserrat
Non-Latin Caribbean	Jamaica

Non-Latin Caribbean	Grenada
Non-Latin Caribbean	Dominica
Non-Latin Caribbean	Curacao
Non-Latin Caribbean	Cayman Islands
Non-Latin Caribbean	British Virgin Islands
Non-Latin Caribbean	Bonaire Saint Eustatius and Saba
Non-Latin Caribbean	Barbados
Non-Latin Caribbean	Bahamas
Non-Latin Caribbean	Aruba
Non-Latin Caribbean	Antigua and Barbuda
Non-Latin Caribbean	Anguilla

Table S2. BSSVS results for CHIKV

FROM	TO	BAYES_FACTOR	POSTERIOR PROBABILITY
Americas	Center-north	86.20856163	0.922105978
Americas	Center-south	0.191914894	0.025676635
Americas	East	0.743297742	0.092614881
Americas	Northeast	0.868885317	0.106595068
Americas	Northwest	0.177265425	0.023763232
Americas	Southeast	1.062599441	0.127333889
Americas	Southwest	5236.520985	0.99861124
Americas	West	15.76167184	0.683979878
Center-north	Center-south	1.060829123	0.127148721
Center-north	East	1.07975133	0.129123847
Center-north	Northeast	1.438552475	0.164953862
Center-north	Northwest	1.671856523	0.186711107
Center-north	Southeast	0.989757074	0.119649415
Center-north	Southwest	1.039937246	0.124957566
Center-north	West	1.07708521	0.128846094
Center-south	East	0.816846997	0.100854859
Center-south	Northeast	8.62935726	0.542326328

Center-south	Northwest	12.1567072	0.625374194
Center-south	Southeast	0.796051295	0.098540259
Center-south	Southwest	0.725594228	0.090608894
Center-south	West	0.743024787	0.09258402
East	Northeast	2.684716558	0.269357776
East	Northwest	7.099083565	0.493627133
East	Southeast	1.128933359	0.134215968
East	Southwest	0.940751488	0.114402987
East	West	0.880728587	0.107891245
Northeast	Northwest	4.998184013	0.406999352
Northeast	Southeast	2.498801692	0.255470173
Northeast	Southwest	1.712473846	0.190383606
Northeast	West	0.956254924	0.1160695
Northwest	Southeast	1.257392275	0.147239453
Northwest	Southwest	1.40068558	0.161312224
Northwest	West	1.255847284	0.147085146
Southeast	Southwest	2.554663903	0.259698176
Southeast	West	1.126535435	0.133969077
Southwest	West	193.5441261	0.963737926
Center-north	Americas	1.603408292	0.180446255
Center-south	Americas	21.97990965	0.751134154
East	Americas	1.081826155	0.129339876
Northeast	Americas	1.553499315	0.175817054
Northwest	Americas	2.00818265	0.216152825
Southeast	Americas	1.362512639	0.157608863
Southwest	Americas	1.233353924	0.144832269
West	Americas	0.939891896	0.114310403
Center-south	Center-north	17.4214616	0.70521248
East	Center-north	0.880728587	0.107891245
Northeast	Center-north	1.367900054	0.158133506
Northwest	Center-north	1.733093224	0.192235287
Southeast	Center-north	1.121743685	0.133475295

Southwest	Center-north	1.464737596	0.167453631
West	Center-north	0.795221687	0.098447675
East	Center-south	0.947634694	0.11514366
Northeast	Center-south	1.27318317	0.148813381
Northwest	Center-south	1.448232368	0.165879702
Southeast	Center-south	1.248439107	0.146344474
Southwest	Center-south	10.16593938	0.582631238
West	Center-south	26.55345456	0.784773015

Table S2. BSSVS results for CHIKV (continued)

FROM	TO	BAYES_FACTOR	POSTERIOR PROBABILITY
Northeast	East	3.538504216	0.327006759
Northwest	East	1.238581511	0.145356911
Southeast	East	4.83050106	0.398790235
Southwest	East	7.167776439	0.496034318
West	East	38.72485751	0.841712187
Northwest	Northeast	1.565426269	0.176928062
Southeast	Northeast	5.627732872	0.435916427
Southwest	Northeast	15.5321893	0.68080116
West	Northeast	0.824916734	0.101749838
Southeast	Northwest	1.907953351	0.207604234
Southwest	Northwest	2.099045763	0.223744715
West	Northwest	0.754778625	0.093911058
Southwest	Southeast	32.14470707	0.81529488
West	Southeast	2.949647346	0.288275777
West	Southwest	0.916750669	0.111810635

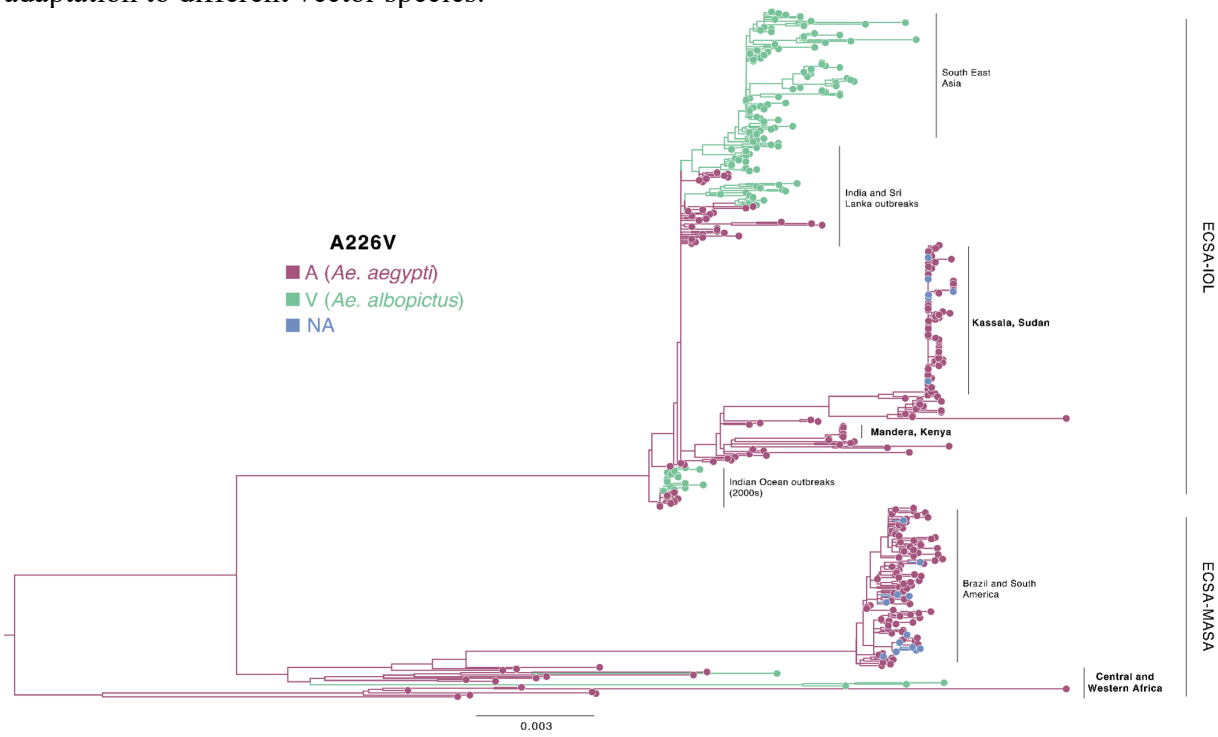
Table S3. BSSVS results for DENV-2

FROM	TO	BAYES_FACTOR	POSTERIOR PROBABILITY
Belize	Center-south	0.967626508	0.133524069
Belize	East	0.953020166	0.131774117
Belize	Guatemala	0.957650717	0.132329658
Belize	Nicaragua	1.70136634	0.213188523
Belize	Southeast	0.930185929	0.129024194
Belize	Southwest	1.025702329	0.140412766
Belize	West	0.942161662	0.130468598
Center-south	East	1.124996057	0.151940224
Center-south	Guatemala	1.150059576	0.154801256
Center-south	Nicaragua	1.410877431	0.183467126
Center-south	Southeast	1.105886879	0.14974584
Center-south	Southwest	1.234011655	0.164245438
Center-south	West	1.001469487	0.137551735
East	Guatemala	1.130092361	0.152523541
East	Nicaragua	1.293408656	0.170800811
East	Southeast	1.210612198	0.161634399
East	Southwest	1.173818724	0.157495625
East	West	1.064144217	0.144912641
Guatemala	Nicaragua	1.391307134	0.18138385
Guatemala	Southeast	1.126936679	0.15216244
Guatemala	Southwest	1.159593991	0.155884559
Guatemala	West	1.146886867	0.154440154
Nicaragua	Southeast	0.338795477	0.051193022
Nicaragua	Southwest	1124.009198	0.994444599
Nicaragua	West	0.263140384	0.040221105
Southeast	Southwest	1.257557779	0.166856476
Southeast	West	1.201194161	0.160578873
Southwest	West	2929.534829	0.997861171
Center-south	Belize	1.714064821	0.214438488
East	Belize	2.004936908	0.242021055
Guatemala	Belize	2.192768344	0.258826144
Nicaragua	Belize	4.066210204	0.393044638
Southeast	Belize	2.42268372	0.278408933
Southwest	Belize	3.166627625	0.335240688
West	Belize	1.245766298	0.165550957
East	Center-south	1.082558193	0.147051471
Guatemala	Center-south	1.105404382	0.149690286
Nicaragua	Center-south	0.251233526	0.038471154
Southeast	Center-south	1.132521639	0.152801311
Southwest	Center-south	0.754762935	0.107302575
West	Center-south	83.24858871	0.929863059

Guatemala	East	1.233262606	0.164162107
Nicaragua	East	0.27129225	0.041415516
Southeast	East	2.168390841	0.256687314
Southwest	East	56.08153389	0.899308353
West	East	0.749951141	0.106691481
Nicaragua	Guatemala	0.823495905	0.115941224
Southeast	Guatemala	1.282010601	0.169550846
Southwest	Guatemala	134.6545105	0.955445682
West	Guatemala	0.753668767	0.10716369
Southeast	Nicaragua	1.480033803	0.190744702
Southwest	Nicaragua	1.196246807	0.160023333
West	Nicaragua	0.800140535	0.113024638
Southwest	Southeast	1270.882832	0.99508347
West	Southeast	0.713638796	0.102052721
West	Southwest	0.753231195	0.107108136

Appendix 6 Supplementary figure from Chapter 6

CHIKV ECSA phylogenetic tree showing amino acid identities associated with viral adaptation to different vector species.



Notes: The amino acid identity at site 226 of the E1 protein is associated with viral adaptation to different vector species. Viruses containing the A226 variant infect Ae. aegypti more efficiently, while viruses containing the V226 variant more efficiently infect Ae. albopictus. The evolution of this trait is mapped (through a parsimony reconstruction) on the tree in different colours.