PROCEEDINGS B

rspb.royalsocietypublishing.org

Review

Cite this article: Pybus OG, Tatem AJ, Lemey P. 2015 Virus evolution and transmission in an ever more connected world. *Proc. R. Soc. B* **282**: 20142878.
http://dx.doi.org/10.1098/rspb.2014.2878

Received: 21 November 2014
Accepted: 15 April 2015

Subject Areas:
health and disease and epidemiology, evolution, microbiology

Keywords:
virus, epidemiology, geography, evolution, phylogenetics, transmission

Author for correspondence:
Oliver G. Pybus
e-mail: oliver.pybus@zoo.ox.ac.uk

One contribution to the special feature 'Evolution and genetics in medicine' Guest edited by Roy Anderson and Brian Spratt.

An invited review to commemorate 350 years of scientific publishing at the Royal Society.

THE ROYAL SOCIETY PUBLISHING

Virus evolution and transmission in an ever more connected world

Oliver G. Pybus[1], Andrew J. Tatem[2,3,4] and Philippe Lemey[5]

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
[2]Department of Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[3]Fogarty International Center, National Institutes of Health, Bethesda, MA, USA
[4]Flowminder Foundation, Stockholm, Sweden
[5]Department of Microbiology and Immunology, Rega Institute, KU Leuven—University of Leuven, Leuven, Belgium

The frequency and global impact of infectious disease outbreaks, particularly those caused by emerging viruses, demonstrate the need for a better understanding of how spatial ecology and pathogen evolution jointly shape epidemic dynamics. Advances in computational techniques and the increasing availability of genetic and geospatial data are helping to address this problem, particularly when both information sources are combined. Here, we review research at the intersection of evolutionary biology, human geography and epidemiology that is working towards an integrated view of spatial incidence, host mobility and viral genetic diversity. We first discuss how empirical studies have combined viral spatial and genetic data, focusing particularly on the contribution of evolutionary analyses to epidemiology and disease control. Second, we explore the interplay between virus evolution and global dispersal in more depth for two pathogens: human influenza A virus and chikungunya virus. We discuss the opportunities for future research arising from new analyses of human transportation and trade networks, as well as the associated challenges in accessing and sharing relevant spatial and genetic data.

1. Introduction

The consequences of international trade and travel for the dynamics of infectious disease are appreciated by researchers and the general public alike. In a highly mobile world, with over half a million travellers in the air at any one moment, viruses have more opportunities than ever before to disseminate globally. Growth in the reach, volume and speed of human mobility over the past century has connected pathogens with new and growing host populations, and contributed to a boom in emerging and re-emerging epidemics [1,2].

The increasing connectivity of our world affects transmission in many ways. Greater mobility, through business travel, tourism and labour movement, leads to more pathogen introductions, while social and ecological changes in recipient locations may raise the likelihood that introductions will become entrenched rather than die out. The establishment of new travel routes between previously unconnected locations also contributes. For example, direct air travel between South America, Africa and southeast Asia now links tropical continental regions, where infectious disease burdens are higher and year-round transmission is more common. Further, the increasing volume of global trade through shipping and air freight can spread contaminated goods or introduce disease vectors such as mosquitoes to new locations through accidental carriage (e.g. [3]).

Despite the importance of geography for infectious disease epidemiology, the effects of global mobility upon the genetic diversity and molecular evolution of pathogens are under-appreciated and only beginning to be understood; indeed, a recent monograph on the spatial epidemiology of

infectious disease makes no reference to pathogen genetic variation [4]. Patterns of host mobility may be particularly important for RNA viruses, the infections on which we focus here. Because many viruses do not survive for long outside the environment of their host, close proximity of hosts (or of hosts and vectors) is often necessary for transmission. Further, because rates of RNA virus mutation and evolution are high, their genomes can accrue genetic differences while being spatially disseminated during an individual outbreak. The evolutionary and spatial dynamics of these pathogens are therefore linked and reciprocally influence each other [5,6]. This fundamental principle has several important consequences. First, genetic differences among viruses sampled from diverse locations will contain information about the spatial processes that gave rise to the virus's geographical distribution. The abundance of viral gene sequences and advances in analytical methods have increased our ability to infer these processes and track viral spread [6]. Second, rapidly evolving viruses are capable of adapting swiftly to the novel environments they encounter as they spread geographically [7], with the potential to alter, for example, vector specificity or sensitivity to drugs or immune responses. Third, spatial sampling provides a common frame of reference whereby virus evolution and migration can be integrated with epidemiological data, or with environmental measurements such as humidity or land use. Integration of geographical data with genetic analysis promises to provide a fuller understanding of the origins, dispersal and dynamics of evolving pathogens [8].

In this article, we explore each of these themes. We first review how spatial and genetic data are combined in empirical studies of viral transmission. Later we discuss in depth two human pathogens, influenza A virus (IAV) and chikungunya virus (CHIKV), whose global dynamics depend critically on the reciprocal interplay between virus evolution, spatial ecology and host mobility.

2. Methods for combining viral spatial and genetic data

Since the contemporary spatial distribution of a fast-evolving virus is the result of interacting ecological and evolutionary processes, consideration of spatial incidence or genetic data in isolation may provide only partial insight into the underlying transmission dynamics [5,9]. Consequently, there is considerable interest in the development new analytical methods, formal and informal, that combine both sources of information.

Several trends in technology and data availability over the last decade have spurred innovation in this area. The advent of cheap, mobile global positioning systems and their widespread adoption in disease surveys has revolutionized the geospatial recording and analysis of infectious disease incidence and prevalence, especially when combined with geographical information systems (GIS) and pervasive electronic communication [10]. Further, a wide range of data (e.g. high-resolution satellite images) that depict environmental, infrastructural and socio-economic variables that may determine disease dynamics are now available. Statistical models have been developed to exploit the relationships between these variables and geo-located disease data, and to predict the spatial distribution of infectious diseases (e.g. [11,12]). Of particular relevance to viruses are new insights into human mobility, generated by the analysis of datasets that describe global air travel passenger numbers [13,14], movements of marked banknotes [15] and anonymized mobile phone call records [16]. The latter have the potential to untangle human mobility in unprecedented detail and have been used to track population mobility following disasters [17], predict infectious disease dynamics [18] and plan disease elimination strategies [19]. At the same time as this progress in disease geography, viral gene sequences have greatly increased in abundance and length, in large part due to the adoption by virologists of next-generation sequencing technologies [20] that typically generate whole viral genomes rather than sub-genomic sequence fragments. Reported pathogen genomes are now more likely to be annotated with locations and dates of sampling, and for the most intensively studied species, such as HIV-1 and influenza, more than 100 000 virus sequences are publicly available.

The term 'phylogeography' is commonly applied to studies that use evolutionary trees to combine genetic data with spatial information [21]. Other statistical methods for examining the spatial distribution of genetic variation do not explicitly use phylogenies and are better described as 'spatial genetics' (reviewed in [22]), while some genealogical approaches to population genetics combine aspects of both approaches (e.g. [23]). Phylogenetic methods are commonly applied to emerging viral epidemics, partly because the rapid evolution of such pathogens can create sufficient genetic variation for analysis at the level of individual infections, even during the early stages of an outbreak, and also because alternative population genetic approaches typically assume that mutation is negligible or that the processes of genetic drift and migration are in equilibrium [24]. The latter were developed with animal or plant populations in mind and may not adequately represent the idiosyncratic and dynamic dispersal histories that characterize ecological invasions [25,26]. Further, a single evolutionary tree (with associated estimation uncertainty) is often sufficient to represent the shared ancestry of all sites in a RNA virus sequence, owing to the absence or low rate of recombination within them.

Methods that attempt to combine viral genetic and geographical information will be worthwhile only if the spatial epidemiology of the pathogen population is recorded in its genome sequences. The degree to which that occurs for the pathogen in question will depend on its relative rates of spatial movement and molecular evolution. A pair of typical RNA virus genomes will diverge genetically from each other on average at a rate of 1–20 nucleotide changes per year (assuming 10^{-3}–10^{-4} substitutions per site per year and a genome 10 000–20 000 nucleotides long [27]). Hence, to a very rough approximation, analyses of viral genomes are unlikely to contain a reliable record of spatial epidemiological trends that occur on time scales faster than a fortnight. It is therefore unsurprising that many studies focus on global or regional patterns, observed over a time scale of several years or decades. Transmission dynamics over short time scales can sometimes be partially resolved by augmenting viral gene sequences with epidemiological incidence data (e.g. [28,29]). It is also possible for virus sequences, particularly those limited to the antigenic regions of capsid or envelope proteins, to evolve too quickly relative to the rate of geographical spread, in which case phylogeographic

information is lost due to the mutational 'saturation' of informative sites in viral genes (e.g. [30]). In general, the rate of pathogen molecular evolution will determine the time scale of the spatial processes that can be reliably inferred; for example, movement of influenza virus can, under the best circumstances, be pinpointed from whole-genome sequences to within a few weeks, whereas geographical trends in the diversity of much slower-evolving *Helicobacter pylori* genes reveal the global spread of the bacterium over more than 50 000 years [31].

Several of the most popular phylogeographic methods for reconstructing epidemic spatial spread from genetic data (e.g. [26,32–34]) treat the location information assigned to each sequence as a discrete or continuous trait, and represent movement as change in that trait along sampled lineages, using stochastic models that are uncoupled from the processes of molecular evolution. The focus is therefore on the locations and ages of sampled lineages rather than on underlying population genetic processes of selection, genetic drift and migration, an approach that may be viewed philosophically as either a strength or a weakness, depending on one's perspective and interests [21,35]. This 'trait evolution' approach to phylogeography facilitates the inference of the locations of common ancestors in an epidemic and can be practically applied to rapidly evolving pathogens with complex spatial dynamics [34]. Further, the inferred changes in location on a phylogeny are statistically independent observations, whereas the sample locations themselves are correlated due to their shared ancestry.

However, it is not always fully recognized that the estimated locations of ancestors can be highly uncertain, particularly those that are only distantly related to the sampled cases. Consequently, viral phylogeography is far more informative when applied to datasets that contain genetic sequences sampled sequentially through time, and which include genomes situated close to the root of the sample phylogeny. A second under-appreciated aspect of phylogeographic analysis is the importance of sample composition [36]. Although a highly detailed spatio-temporal record may not be required to address every important question about pathogen spread, the accuracy with which gene sequences can capture key patterns will depend on the representativeness of sampling. If samples from key locations or regions are absent or rare then virus movement will be underestimated and the inferred locations of ancestors may be biased towards locations that are over-represented in the sample. As a result, phylogeographic results should be interpreted carefully, combined with other sources of epidemiological information and statistically validated whenever possible.

3. Integration of viral spatial and genetic data in practice

The simplest way to combine viral spatial and genetic data is through the mapping of infections attributable to different viral strains. This creates a link to genetic variation because RNA viruses are classified into genotypes and subtypes using analysis of their gene sequences. In recent years, the global geographical distribution of strains of HIV-1 [37], dengue virus [11] and hepatitis B and C viruses [38,39] have been characterized in this way. Despite being primarily descriptive, such studies can be useful in public health planning. For example, severe disease following dengue virus infection is more common in regions where two or more viral serotypes co-circulate, and the success rate of drug treatment for hepatitis C virus infection varies significantly among viral genotypes.

Evolutionary analysis of viral genes can be used to validate the putative source of an emerging viral outbreak that has been identified through epidemiological surveillance and contact tracing. For example, the proposed index case of the 2007 outbreak of CHIKV in northeast Italy had hosted a relative from Kerala, India (where the virus was epidemic), and phylogenetic analysis of virus E1 gene sequences from the Italian outbreak showed it to be very closely related to strains previously isolated in India [40]. Independent testing of an outbreak's source using viral genetics is especially valuable when surveillance data is uncertain or absent, and may become commonplace as viral genome sequencing becomes routine in clinical diagnosis. It is therefore important that public health agencies recruit and retain expertise in the evolutionary analysis of pathogen genetic variation.

In addition to its confirmatory role, analysis of virus genomes can answer questions of relevance to infectious disease control that cannot be addressed using incidence reports alone. For example, viral phylogenies can indicate if an outbreak in a specific region is the result of a single introduction followed by onward transmission within the host population of that region, or is composed of multiple independent chains of transmission, each initiated by a separate introduction from elsewhere or from a zoonotic reservoir species. For example, analysis of viral genomes from the Ebola epidemic in west Africa that began in Guinea in early 2014 indicated that it developed from a single introduction from the virus's reservoir in central Africa, and that the epidemic in Sierra Leone arose from the transmission of two distinct viral lineages from Guinea [41]. By contrast, phylogenetic investigation of the HIV-1 subtype B epidemic in the UK showed that it comprised hundreds of independent viral introductions from other countries, at least six of which established large and persistent chains of transmission in the UK [42]. Epidemiological differences among observed transmission chains can help to focus epidemic control efforts more efficiently on specific populations or risk groups. Further epidemiological insights can be obtained by using evolutionary 'molecular clock' models, which place viral phylogenies on a real time scale of months and years [8], and enable estimation of the age of the most recent common ancestor (MRCA) of transmission chains in different locations. It is not always appreciated that the MRCA of an outbreak does not necessarily represent the same infected individual as the index case; the former can be more recent (but never older) than the latter. Despite this condition, estimated MRCA ages are sometimes weeks to years earlier than reported dates of virus discovery. Thus, this difference indicates a 'time lag' of epidemiological surveillance, the duration of which might be used to evaluate the efficiency and timeliness of systems of pathogen detection and notification.

If transmission is predominantly local and movement unimpeded by geographical barriers then the genetic and geographical distances among sampled infections are expected to be positively correlated. This principle, known as isolation by distance [24], forms a simple yet direct link between genetic and spatial information, and represents an important null hypothesis in spatial genetics. Strong correlations may be observed for viruses that disperse gradually,

such as rice yellow mottle virus during its spread across sub-Saharan Africa [43]. However, patterns of isolation by distance can be swiftly lost if landscape features affect the dynamics of spread. A study of Zaire ebolavirus in central Africa suggested that the epizootic underwent an abrupt change in direction at a major biogeographic river barrier [44]. Rerouting geographical distances between virus sequences through this 'pivot point' led to much stronger correlations of genetic and geographical distances than when straight line distances were used [44]. Evidence for isolation by distance may be also eroded by high rates of host movement (a topic discussed later in the context of influenza viruses). The Ebola study, and others (e.g. [45]), illustrate the importance of using the locations of ancestral infections when reconstructing the geographical distance travelled by the chain of transmission that connects two sampled cases, especially when dissemination is not uniform in space. As discussed above, ancestral locations are typically inferred using one of the 'trait evolution' phylogeography methods.

Highly heterogeneous dispersal may be a common feature of all ecological invasions [46]. This variation has been accommodated in phylogeographic analysis using 'relaxed random walk' models that allow dispersal rates to vary significantly among phylogeny branches [34]. Application of this approach to the West Nile virus invasion of North America that began in New York in 1999 revealed that the epidemic was driven by a heterogeneous mix of local transmission and rare, long-range viral movements that probably represent seasonal migration of birds, the natural hosts of the virus [47]. An important consequence of such approaches is that each phylogeny branch becomes an independent observation of viral translocation, conditional on the data. This enables spatial epidemiological parameters, such as the epidemic diffusion coefficient and wavefront velocity, to be readily estimated from viral genome sequences alone [47].

A key goal of viral phylogeography is to help predict future pathogen spread by indicating those social or environmental factors that are associated with virus movement. This is often achieved by qualitative comparison of virus genetic diversity or dispersal history with geographical data. For example, the early spread of HIV-1 in east Africa was explored by combining phylogenetic analyses with regional data on road network architecture and population density, obtained using GIS techniques [48]. More recent phylogeographic studies have formalized this approach by parametrizing location exchange rates as a function of different potential causal factors, so that the effects of these drivers of spatial spread can be quantified and tested using genetic data [49]. Crucially, this enables virus genomes and host mobility data to be combined in a single statistical framework. Retrospective application of this technique to the 2009 influenza A pandemic demonstrated that combining both data sources predicted the global dissemination of the pandemic better than either alone [49].

4. Case study: influenza virus

In addition to generating information essential for vaccine design, the global surveillance of influenza viruses has resulted in an unparalleled collection of virus genome sequences sampled through space and time, providing an opportunity to explore the processes that underpin the global dynamics of this important pathogen [50]. Although human influenza is primarily transmitted in household and community settings, epidemics of IAVs in temperate climates are seasonal and experience strong genetic bottlenecks, implying that transmission in these locations is typically not sustained and that epidemics are re-established by the importation of viral lineages from populations in which transmission is more persistent [51–53]. This so-called 'source–sink' model of global IAV circulation has been investigated in detail for the H3N2 subtype of IAV (figure 1), a dominant strain of human influenza since its emergence in 1968. Various studies have used phylogeographic and population genetic methods to infer the location through time of the 'source' population of H3N2 influenza, and most conclude that it resides primarily in east or southeast Asia [49,52,54] (figure 1). However, temperate regions, particularly the USA, may also contribute as a source [55], and there is evidence for viral gene flow into Asia from elsewhere [56], suggesting that the migration dynamics of H3N2 influenza are more complex than those represented by a simple source–sink model. Differences among these studies may however be attributable to variation in analysis methodology and sequence sampling strategy; seasonal fluctuations in sampling and the comparative under-sampling of IAV from south Asia, Africa and Latin America means that conclusions should be interpreted carefully [36]. Nevertheless, all analyses implicate global mobility as a driver of worldwide human influenza virus dispersal (figure 1); air passenger flux is a considerably better predictor of the movement of IAV lineages among locations than geographical distance [13,49]. Thus, the spatial genetics of human influenza, and possibly of other pathogens, may be better characterized by 'proximity by mobility' than by the traditional notion of 'isolation by distance'.

The emergence of pandemic H1N1 (pH1N1) influenza in 2009 was the first influenza pandemic in the post-genomic era. Genetic analysis of the pandemic in its early stages was aided by pre-planned and intensive virus sequencing in some countries, and by the immediate and open sharing of the resulting data through online databases. Consequently, the molecular epidemiology of the virus could be tracked in 'real time' as the epidemic unfolded [57,58]. This included phylogeographic analyses that studied the global dispersal of the virus during its establishment phase [59,60], which followed patterns of international air travel [13,61]. The intensive sampling of virus sequences during the pandemic enabled the molecular epidemiology of IAV to be scrutinized at such a high resolution that the importation, extinction and persistence of individual transmission chains in specific locations could be observed (e.g. [62–64]). Comparisons among countries of the dynamics of transmission chains may provide useful insights. For example, only two of many pH1N1 lineages that were introduced to the UK at the start of the pandemic were detected there six months later [64], while a pair of pH1N1 transmission chains appear to have persisted in west Africa for almost 2 years [65]. The latter observation seems to be at odds with the extensive spatial mixing of IAV imposed by air travel, but west Africa is connected comparatively weakly within the global air transportation network [66] and influenza persistence might be facilitated there by climatic variability that can generate temporal overlap among epidemics in neighbouring regions [65], as has been previously suggested for IAV persistence in southeast Asia [52].

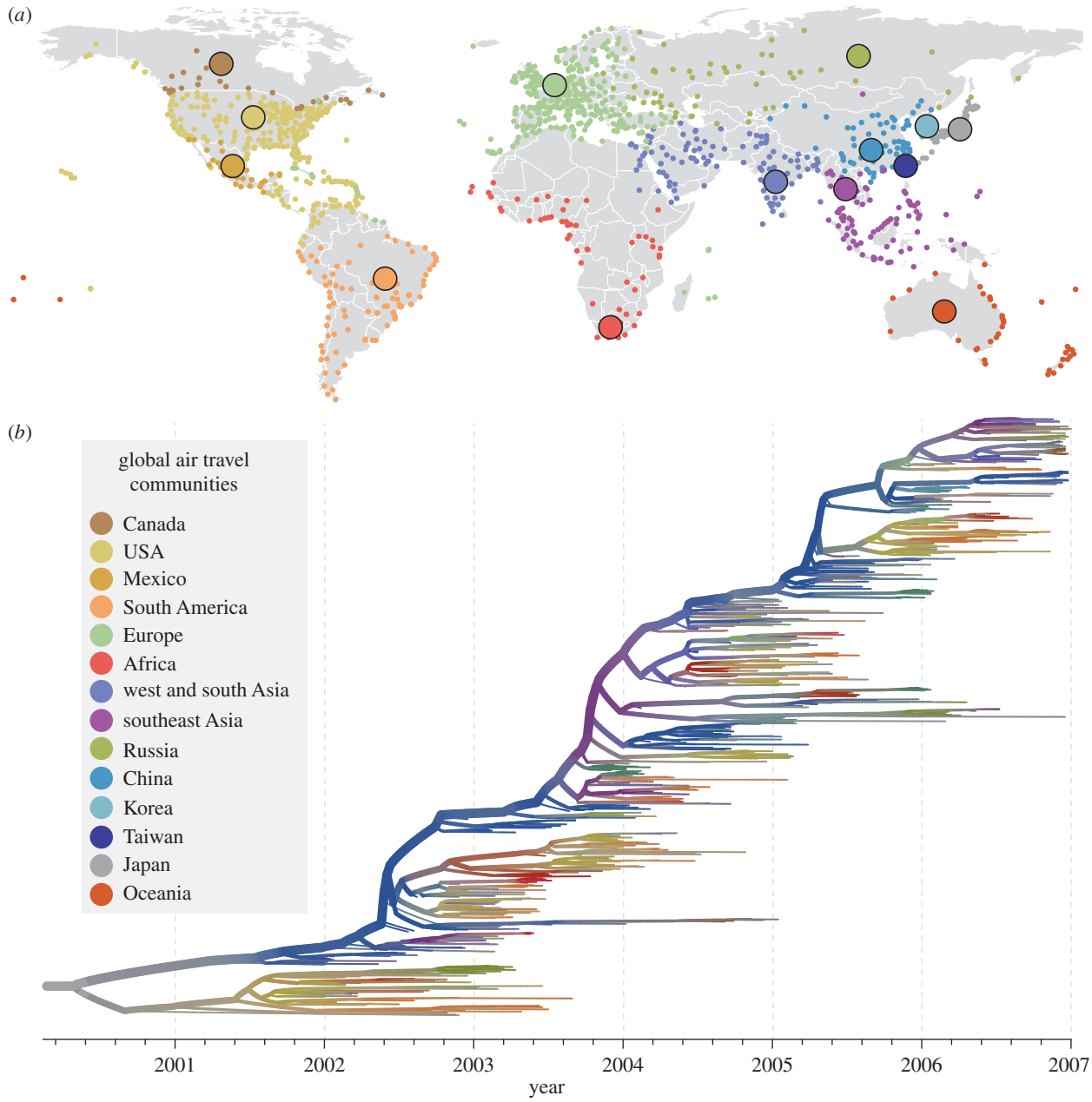

Figure 1. (*a*) The modular structure of global air travel. Airports (small dots) can be grouped into 14 communities (colours; inset) such that there is high connectivity within communities but low connectivity among them (hence French Guiana belongs to the European, not South American, community). Larger circles indicate the approximate geographical centre of each community. (*b*) A phylogeny of the H3N2 subtype of human IAV, estimated from more than 1000 virus haemagglutinin gene sequences that were sampled worldwide between 2002 and 2007. A molecular clock model was used, hence phylogeny branches represent time (time scale shown below the tree). The thickness of each branch is proportional to its number of descendent tips (up to a maximum thickness) and indicates lineage persistence. Each phylogeny branch is coloured according to its most probable location, which was inferred using a phylogeographic model that takes into account the global air travel network. The thicker, uppermost lineage represents the most persistent lineage of H3N2 influenza, which, for most years, is estimated to be located in southeast or east Asia. Figure adapted from Lemey *et al.* [49].

Local persistence of transmission chains also raises questions about the mobility processes that drive IAV spread at sub-national scales. Mathematical analyses of mortality and physician visit statistics have suggested different drivers for the spread of seasonal [67,68] and pandemic [69] influenza across the continental US. These studies variably emphasized the relative importance of workplace commuting [68], domestic airline travel [67] and school opening dates [69]. As an independent source of information about transmission, viral genetic data could help to resolve this problem. However, it is possible that sub-genomic influenza haemagglutinin gene sequences do not contain sufficient information to answer fine-scaled questions about viral dispersal over very short time scales. Instead, complete viral genome sequences will probably be needed to achieve the phylogeographic resolution required.

The spatial dynamics of influenza are also critical in assessing the evolution of anti-viral drug resistance. The global cycling of IAV lineages and low levels of local persistence mean that resistance mutations can spread worldwide, and can quickly erode any association at the local level between rates of drug usage and viral resistance. Recent examples of anti-viral drug resistance evolution include the rapid spread in oseltamivir resistance in seasonal H1N1 influenza from 2007 to 2009 and the global rise of adamantane-resistant H3N2 influenza during 2003–2006. An investigation of the former that used a stochastic model of international air

travel concluded that the oseltamivir-resistant strain rose to global dominance because it exhibited a transmission advantage in untreated hosts, probably conferred by genetic hitchhiking [70]. Phylogeographic analysis of adamantane resistance in A/H3N2 IAV has shown that resistance evolved independently 11 times over 10 years [71], yet most of the resistant viruses found were descended from a single resistant lineage that was first detected in southeast Asia in 2003, before later spreading worldwide, consistent with the above-mentioned 'source–sink' model of global IAV circulation.

5. Case study: chikungunya virus

CHIKV is a mosquito-borne alphavirus that, while rarely fatal, causes a debilitating fever and sometimes persistent arthralgia, so is of some public health concern. In the 50 years that followed the virus's discovery in 1952 in Tanzania, sporadic outbreaks were reported in central, west and east Africa, and in south and southeast Asia [72]. However, the last decade has seen an increase in the geographical range of CHIKV. Starting from east Africa in 2004, CHIKV epidemics were reported increasingly eastwards, first on Indian Ocean islands (Comoros, Reunion, Seychelles and Mauritius) in 2005–2006, then in India and Sri Lanka in 2006–2007 [73]. Numerous countries in temperate regions have reported imported cases, one of which, in Italy, caused an autochthonous epidemic [40]. However, it is only within the last 18 months that CHIKV has finally become established in the New World. More than 750 000 suspected cases in the Americas have been reported since the detection of CHIKV on the Caribbean island of Saint Martin in December 2013, and several mathematical models that use data on human mobility and vector distributions have already been developed to predict further spread of the virus in the Americas (e.g. [74]).

The worldwide expansion of CHIKV has left a clear footprint in the genomic diversity of the virus, despite the fact that its rate of molecular evolution is somewhat slower than that of viruses like influenza and HIV [75]. Phylogeographic analysis of CHIKV genomes (figure 2) shows that two virus lineages (the 'Asian genotype' and the 'Indian Ocean lineage') were responsible for the recent expansion of its geographical range. The Asian genotype, first detected in India in the 1960s, is the strain that has recently emerged in the Caribbean and appears to have reached there via southeast Asia and Micronesia. By contrast, the Indian Ocean lineage was responsible for the significant epidemics in south Asia from 2005 onwards (figure 2) [75].

Multiple genetic and ecological factors are thought to have contributed to the global emergence of CHIKV. The two mosquito species principally responsible for human CHIKV transmission are *Aedes aegypti* and *Ae. albopictus*. The collapse of *Ae. aegypti* elimination efforts in the Americas [76] and growing urbanization in the tropics and sub-tropics has provided suitable habitats for this primary vector. Additionally, the globalization of trade in used tyres during the 1980s and 1990s enabled the secondary vector *Ae. albopictus* to expand its range from southeast Asia to large parts of the rest of the world [3]. Further, greater human travel between Africa, Asia and the Americas has increased interchange between locations where *Aedes* mosquitoes are prevalent, including at times of the year when the vectors are highly active in both places [73,77].

In addition to these ecological factors, there is strong evidence that, as the geographical range of CHIKV expanded, the virus evolved and adapted to local variation in the distribution of vector species. Specifically, a single amino acid change (A226V) in the viral E1 protein has been shown to increase transmission and infectivity in *Ae. albopictus* mosquitoes [78]. This mutation arose multiple times within the Indian Ocean lineage, usually in locations where *Ae. albopictus* was the sole or dominant vector species [79], and thus represents a remarkable example of convergent molecular evolution (figure 2). Fortunately, the Asian lineage that has recently emerged in the Americas has, to date, shown no propensity to evolve mutations that elevate transmissibility in *Ae. albopictus* mosquitoes.

6. Discussion

Our understanding and evaluation of the risks of infectious disease spread are being refined by access to growing geographically referenced databases of disease prevalence, detailed satellite-based imagery and unprecedented information about patterns of human mobility. Successful integration of these sources of information with viral genetic data will be technically and intellectually challenging, yet holds great promise for our response to emerging viruses.

Recent modelling work indicates that pathogen diffusion becomes highly regular when measured against a so-called 'effective distance' along the relevant mobility or transport network [13]. Conceptually, this requires translating from variable rates of spread through a space defined by geographical distances, to regular diffusion through a space defined by effective distances. The former process is already accommodated by phylogeographic analysis [34] so implementation of the latter should be possible. This work suggests that empirically derived networks of contacts among hosts may constitute a third common frame of reference by which genetic and epidemiological data can be unified, supplementing the temporal and spatial dimensions that are currently used [8]. In future, the concept of effective distances could be extended to epizootic or vector-borne pathogens, for which landscape heterogeneity is more important than human contact networks. Previous work has already shown the possibility of defining 'climatic distances' that account for differences among locations and seasons in their suitability for vector-borne disease transmission [3]. Integrating genetic data in this context will require a melding of phylogeographic and GIS techniques [80,81] in order to detect more subtle deviations from distance-dependent movement than those imposed by human transportation networks.

A significant obstacle to further progress is the availability and expense of some of the most powerful and relevant datasets. For air travel, origin–destination data derived from air ticket sales are available, but are highly expensive for research purposes, and their use may require legal and confidentiality agreements, resulting in a reliance on modelled datasets [14]. Moreover, detailed data on human mobility derived from mobile phone call records often prompt privacy and commercial concerns. Although virus genetic data are usually deposited in publicly accessible databases such as GenBank upon publication of the paper that report them, the delay between sequence generation and publication may prevent the opportunity to undertake real-time molecular

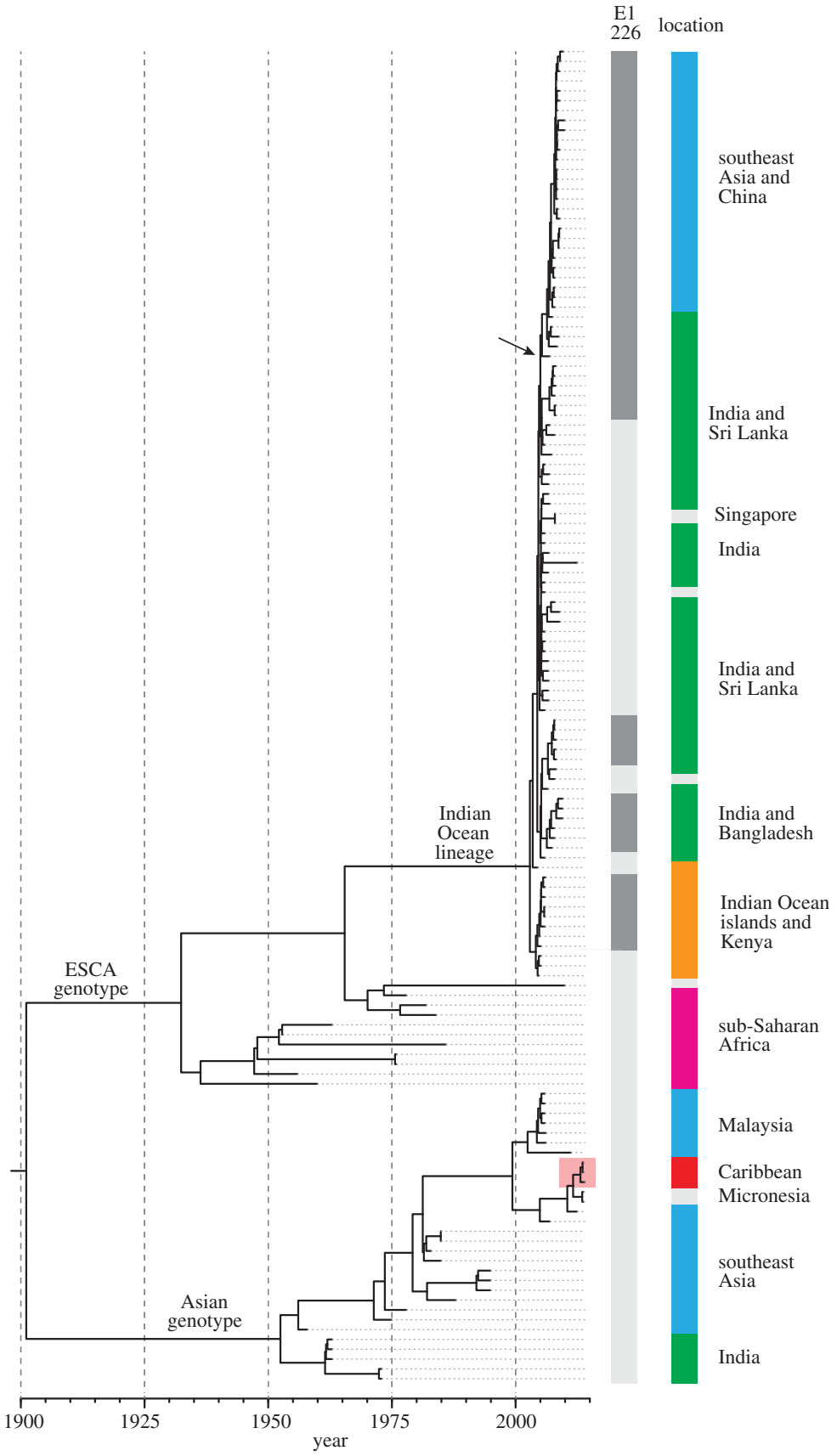

Figure 2. The evolution and global spread of CHIKV. On the left is a phylogeny of CHIKV, estimated from whole genomes of viruses sampled from the 1960s to the present day. Major CHIKV lineages are denoted (the west Africa genotype is not shown). The first vertical bar on the right indicates the amino acid present at position 226 in the CHIKV E1 protein (dark grey, valine; light grey, alanine). A change to valine at this site confers increased transmissibility of the virus in *Aedes albopictus* mosquitoes (see main text). The second vertical bar indicates the geographical location of the viruses (green, south Asia; blue, southeast Asia or China; orange, east Africa or Indian Ocean islands; purple, sub-Saharan Africa; red, Americas; grey, other locations). For returning travellers, the location of infection (not the location of detection) is shown. An arrow indicates the strain that caused an outbreak in Italy in 2007 (see main text). A red box indicates the lineage responsible for the recent emergence of CHIKV in the Americas.

epidemiology during an outbreak. Further, genetic data obtained by surveillance efforts may be reported without essential epidemiological information, such as the date and location of sampling, or may never be published at all, for reasons of commerce, politics or privacy. The success of GISAID (http://gisaid.org) and other initiatives in facilitating the timely sharing of influenza virus genomes during the 2009 H1N1 pandemic has unfortunately not been repeated in subsequent outbreaks. We strongly support the recent call for an international and inter-disciplinary consensus towards the open sharing and release of pathogen genetic information during epidemics [82].

New outbreaks of infectious disease, especially those caused by viruses, are a common phenomenon in the twenty-first century, and future trends in global mobility and trade seem likely to maintain or even accelerate their rate of appearance. Techniques and data to describe, explain and predict such occurrences can help to measure and mitigate the risks from novel and re-emerging pathogens. Statistical and mathematical models that integrate spatially explicit data on pathogen evolution with information on human movement and environmental variability have much to contribute to epidemic management, as well as deepening our understanding of fundamental evolutionary and ecological processes.

Note added in proof

Since this review was written, the Asian genotype of CHIKV has spread from the Caribbean to Mexico, Brazil and Columbia, and local transmission has been observed in mainland France and Florida, USA. A second CHIKV genotype (ESCA) appears to have been introduced to Brazil from central Africa [83]. In addition, two recent studies have provided further insights into the interplay between human mobility and IAV evolution and transmission. Bozick & Real [84] showed that interstate commuter networks in the USA match the spatial genetic variation of IAV subtype H1N1. Bedford *et al*. [85] reported that age-dependent differences in infection and air travel frequency can explain the distinct evolutionary behaviours of influenza A and B viruses.

Competing interests. We declare we have no competing interests.

Funding. O.G.P. received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 614725-PATHPHYLODYN. A.J.T. is supported by funding from NIH/NIAID (U19AI089674), the Bill and Melinda Gates Foundation (OPP110642749446, 1032350), the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security, Wellcome Trust Sustaining Health Grant, 106866/Z/15/Z, and the Fogarty International Center, National Institutes of Health. P.L. acknowledges funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864, as well as funding, the from Onderzoeksfonds KU Leuven/Research Fund KU Leuven.

Acknowledgements. Many thanks to Nuno Faria for assistance in composing figure 2 and to Trevor Bedford for the original tree visualization in figure 1. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008 Global trends in emerging infectious diseases. *Nature* **451**, 990–993. (doi:10.1038/nature06536)
2. Smith KF, Goldberg M, Rosenthal S, Carlson L, Chen J, Chen C, Ramachandran S. 2014 Global rise in human infectious disease outbreaks. *J. R. Soc. Interface* **11**, 20140950. (doi:10.1098/rsif.2014.0950)
3. Tatem AJ, Hay SI, Rogers DJ. 2006 Global traffic and disease vector dispersal. *Proc. Natl Acad. Sci. USA* **103**, 6242–6247. (doi:10.1073/pnas.0508391103)
4. Sattenspiel L. 2009 *The geographic spread of infectious diseases: models and applications*. Princeton, NJ: Princeton University Press.
5. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)
6. Holmes EC. 2004 The phylogeography of human viruses. *Mol. Ecol.* **13**, 745–756. (doi:10.1046/j.1365-294X.2003.02051.x)
7. Ally D, Wiss VR, Deckert GE, Green D, Roychoudhury P, Wichman HA, Brown CJ, Krone SM. 2014 The impact of spatial structure on viral genomic diversity generated during adaptation to thermal stress. *PLoS ONE* **9**, e88702. (doi:10.1371/journal.pone.0088702)
8. Pybus OG, Rambaut A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)
9. Real LA *et al.* 2005 Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proc. Natl Acad. Sci. USA* **102**, 12 107–12 111. (doi:10.1073/pnas.0500057102)
10. Hay SI *et al.* 2013 Global mapping of infectious disease. *Phil. Trans. R. Soc. B* **368**, 20120250. (doi:10.1098/rstb.2012.0250)
11. Bhatt S *et al.* 2013 The global distribution and burden of dengue. *Nature* **496**, 504–507. (doi:10.1038/nature12060)
12. Gilbert M *et al.* 2014 Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nat. Commun.* **5**, 4116 (doi:10.1038/ncomms5116).
13. Brockmann D, Helbing D. 2013 The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342. (doi:10.1126/science.1245200)
14. Huang Z, Wu X, Garcia AJ, Fik TJ, Tatem AJ. 2013 An open-access modeled passenger flow matrix for the global air network in 2010. *PLoS ONE* **8**, e64317. (doi:10.1371/journal.pone.0064317)
15. Brockmann D, Hufnagel L, Geisel T. 2006 The scaling laws of human travel. *Nature* **439**, 462–465. (doi:10.1038/nature04292)
16. González MC, Hidalgo CA, Barabási AL. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
17. Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J. 2011 Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Med.* **8**, e1001083. (doi:10.1371/journal.pmed.1001083)
18. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. 2012 Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270. (doi:10.1126/science.1223467)
19. Tatem AJ *et al.* 2014 Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malar J.* **13**, 52. (doi:10.1186/1475-2875-13-52)
20. Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N. 2012 Application of next-generation sequencing technologies in virology. *J. Gen. Virol.* **93**, 1853–1868. (doi:10.1099/vir.0.043182-0)
21. Avise JC. 2000 *Phylogeography*. Cambridge, MA: Harvard University Press.
22. Guillot G, Leblois R, Coulon A, Frantz AC. 2009 Statistical methods in spatial genetics. *Mol. Ecol.* **18**, 4734–4756. (doi:10.1111/j.1365-294X.2009.04410.x)
23. Hudson RR. 1991 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

24. Slatkin M. 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279. (doi:10.2307/2410134)
25. Neigel JE, Ball R, Avise JC. 1991 Estimation of single generation migration distances from geographic variation in animal mitochondrial DNA. *Evolution* **45**, 423–432. (doi:10.2307/2409675)
26. Sanmartín I, Van Der Mark P, Ronquist F. 2008 Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *J. Biogeogr* **35**, 428–449. (doi:10.1111/j.1365-2699.2008.01885.x)
27. Duffy S, Shackelton LA, Holmes EC. 2008 Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276. (doi:10.1038/nrg2323)
28. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT. 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* **275**, 887–895. (doi:10.1098/rspb.2007.1442)
29. Ypma RJF, Bataille AMM, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc.* **279**, 444–450. (doi:10.1098/rspb.2011.0913)
30. Coyne KP, Christley RM, Pybus OG, Dawson S, Gaskell RM, Radford AD. 2012 Large scale spatial and temporal genetic diversity of feline calicivirus. *J. Virol.* **86**, 11 356–11 367. (doi:10.1128/JVI.00701-12)
31. Linz B *et al.* 2007 An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918. (doi:10.1038/nature05562)
32. Slatkin M, Maddison WP. 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613.
33. Lemmon AR, Lemmon EM. 2008 A likelihood framework for estimating phylogeographic history on a continuous landscape. *Syst. Biol.* **57**, 544–561. (doi:10.1080/10635150802304761)
34. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010 Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885. (doi:10.1093/molbev/msq067)
35. Nielsen R, Beaumont MA. 2009 Statistical inferences in phylogeography. *Mol. Ecol.* **18**, 1034–1047. (doi:10.1111/j.1365-294X.2008.04059.x)
36. Viboud C, Nelson MI, Tan Y, Holmes EC. 2013 Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. *Phil. Trans. R. Soc. B* **368**, 20120199. (doi:10.1098/rstb.2012.0199)
37. Hemelaar J, Gouws E, Ghys PD, Osmanov S. 2006 Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* **20**, W13–W23. (doi:10.1097/01.aids.0000247564.73009.bc)
38. Shi W, Zhang Z, Ling C, Zheng W, Zhu C, Carr MJ, Higgins DG. 2013 Hepatitis B virus subgenotyping: history, effects of recombination, misclassifications, and corrections. *Infect. Genet. Evol.* **16**, 355–361. (doi:10.1016/j.meegid.2013.03.021)
39. Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, Barnes E. 2015 The global distribution and prevalence of HCV genotypes. *Hepatology* **61**, 77–87. (doi:10.1002/hep.27259)
40. Rezza G *et al.* 2007 Infection with chikungunya virus in Italy: an outbreak in a temperate region. *Lancet* **370**, 1840–1846. (doi:10.1016/S0140-6736(07)61779-6)
41. Gire SK *et al.* 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372. (doi:10.1126/science.1259657)
42. Hué S, Pillay D, Clewley J, Pybus OG. 2005 Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl Acad. Sci. USA* **102**, 4425–4429. (doi:10.1073/pnas.0407534102)
43. Fargette D *et al.* 2004 Inferring the evolutionary history of rice yellow mottle virus from genomic, phylogenetic, and phylogeographic studies. *J. Virol.* **78**, 3252–3261. (doi:10.1128/JVI.78.7.3252-3261.2004)
44. Walsh PD, Biek R, Real LA. 2005 Wave-like spread of Ebola Zaire. *PLoS Biol.* **3**, e371. (doi:10.1371/journal.pbio.0030371)
45. Lam TT *et al.* 2012 Phylodynamics of H5N1 avian influenza virus in Indonesia. *Mol. Ecol.* **21**, 3062–3077. (doi:10.1111/j.1365-294X.2012.05577.x)
46. Melbourne BA, Hastings A. 2009 Highly variable spread rates in replicated biological invasions: fundamental limits to predictability. *Science* **325**, 1536–1539. (doi:10.1126/science.1176138)
47. Pybus OG *et al.* 2012 Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl Acad. Sci. USA* **109**, 15 066–15 071. (doi:10.1073/pnas.1206598109)
48. Gray RR *et al.* 2009 Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* **23**, F9–17. (doi:10.1097/QAD.0b013e32832faf61)
49. Lemey P *et al.* 2014 Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* **10**, e1003932. (doi:10.1371/journal.ppat.1003932)
50. Ferguson NM, Galvani AP, Bush RM. 2003 Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433. (doi:10.1038/nature01509)
51. Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. 2007 Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.* **3**, 1220–1228. (doi:10.1371/journal.ppat.0030131)
52. Russell CA *et al.* 2008 The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320**, 340–346. (doi:10.1126/science.1154137)
53. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008 The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619. (doi:10.1038/nature06945)
54. Chan J, Holmes A, Rabadan R. 2010 Network analysis of global influenza spread. *PLoS Comput. Biol.* **6**, e1001005. (doi:10.1371/journal.pcbi.1001005)
55. Bedford T, Cobey S, Beerli P, Pascual M. 2010 Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog.* **6**, e1000918. (doi:10.1371/journal.ppat.1000918)
56. Bahl J *et al.* 2011 Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl Acad. Sci. USA* **108**, 19 359–19 364. (doi:10.1073/pnas.1109314108)
57. Fraser C *et al.* 2009 Pandemic potential of a novel strain of influenza A (H1N1): early findings. *Science* **324**, 1557–1561. (doi:10.1126/science.1176062)
58. Hedge J, Lycett SJ, Rambaut A. 2013 Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biol. Lett.* **9**, 20130331. (doi:10.1098/rsbl.2013.0331)
59. Jombart T, Eggo RM, Dodd P, Balloux F. 2009 Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic. *PLoS Curr.* **1**, RRN1026. (doi:10.1371/currents.RRN1026)
60. Lemey P, Suchard M, Rambaut A. 2009 Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr.* **1**, RRN1031. (doi:10.1371/currents.RRN1031)
61. Khan K *et al.* 2009 Spread of a novel influenza A (H1N1) virus via global airline transportation. *N. Engl. J. Med.* **361**, 212–214. (doi:10.1056/NEJMc0904559)
62. Shiino T *et al.* 2010 Molecular evolutionary analysis of the influenza A(H1N1)pdm, May–September, 2009: temporal and spatial spreading profile of the viruses in Japan. *PLoS ONE* **5**, e11057. (doi:10.1371/journal.pone.0011057)
63. Nelson MI *et al.* 2011 Phylogeography of the spring and fall waves of the H1N1/09 pandemic influenza virus in the United States. *J. Virol.* **85**, 828–834. (doi:10.1128/JVI.01762-10)
64. Baille GJ *et al.* 2012 Evolutionary dynamics of local pandemic H1N1/09 influenza lineages revealed by whole genome analysis. *J. Vir.* **86**, 11–18. (doi:10.1128/JVI.05347-11)
65. Nelson MI *et al.* 2014 Multiyear persistence of 2 pandemic A/H1N1 influenza virus lineages in West Africa. *J. Infect Dis.* **210**, 121–125. (doi:10.1093/infdis/jiu047)
66. Nzussouo NT *et al.* 2012 Delayed 2009 pandemic influenza A virus subtype H1N1 circulation in West Africa, May 2009-April 2010. *J. Infect Dis.* **206**, 1026. (doi:10.1093/infdis/jis572)
67. Brownstein JS, Wolfe CJ, Mandl KD. 2006 Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Med.* **3**, e401. (doi:10.1371/journal.pmed.0030401)
68. Viboud C, Miller MA, Grenfell BT, Bjornstad ON, Simonsen L. 2006 Air travel and the spread of

influenza: important caveats. *PLoS Med*. **3**, e503; author reply e502. (doi:10.1371/journal.pmed.0030503)

69. Gog JR *et al.* 2014 Spatial Transmission of 2009 Pandemic influenza in the US. *PLoS Comput. Biol*. **10**, e1003635. (doi:10.1371/journal.pcbi.1003635)
70. Chao DL, Bloom JD, Kochin BF, Antia R, Longini IMJr. 2012 The global spread of drug-resistant influenza. *J. R. Soc. Interface* **9**, 648–656. (doi:10.1098/rsif.2011.0427)
71. Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. 2009 The origin and global emergence of adamantane resistant A/H3N2 influenza viruses. *Virology* **388**, 270–278. (doi:10.1016/j.virol.2009.03.026)
72. Powers AM, Logue CH. 2007 Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. *J. Gen. Virol*. **88**, 2363–2377. (doi:10.1099/vir.0.82858-0)
73. Charrel RN, de Lamballerie X, Raoult D. 2007 Chikungunya outbreaks—the globalization of vectorborne diseases. *N. Engl. J. Med*. **356**, 769–771. (doi:10.1056/NEJMp078013)
74. Cauchemez S, Ledrans M, Poletto C, Quenel P, de Valk H, Colizza V, Boëlle PY. 2014 Local and regional spread of chikungunya fever in the Americas. *Euro Surveill.* **19**, 20854. (doi:10.2807/1560-7917.ES2014.19.28.20854)
75. Volk SM *et al.* 2010 Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. *J. Virol*. **84**, 6497–6504. (doi:10.1128/JVI.01603-09)
76. Camargo S. 1967 History of *Aedes aegypti* eradication in the Americas. *Bull. World Health Organ*. **36**, 602–603.
77. Tatem AJ, Huang Z, Das A, Qi Q, Roth J, Qiu Y. 2012 Air travel and vector-borne disease movement. *Parasitology* **139**, 1816–1830. (doi:10.1017/S0031182012000352)
78. 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. (doi:10.1371/journal.ppat.0030201)
79. de Lamballerie X, Leroy E, Charrel RN, Ttsetsarkin K, Higgs S, Gould EA. 2008 Chikungunya virus adapts to tiger mosquito via evolutionary convergence: a sign of things to come? *Virol. J.* **5**, 33. (doi:10.1186/1743-422X-5-33)
80. Kidd DM, Ritchie MG. 2006 Phylogeographic information systems: putting the geography into phylogeography. *J. Biogeogr*. **33**, 1851–1865. (doi:10.1111/j.1365-2699.2006.01574.x)
81. Biek R, Real LA. 2010 The landscape genetics of infectious disease emergence and spread. *Mol. Ecol.* **19**, 3515–3531. (doi:10.1111/j.1365-294X.2010.04679.x)
82. Yozwiak NL, Schaffner SF, Sabeti PC. 2015 Make outbreak research open access. *Nature* **518**, 477–479. (doi:10.1038/518477a)
83. Nunes MR *et al.* 2015 Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Medicine* **13**, 102.
84. Bozick BA, Real LA. 2015 The role of human transportation networks in mediating the genetic structure of seasonal influenza in the United States. *PLoS Pathogens* **11**, e1004898.
85. Bedford T *et al.* 2015 Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 271–220. (doi:10.1038/nature14460)