MODELLING

Evolutionary analysis of the dynamics of viral infectious disease

Oliver G. Pybus and Andrew Rambaut‡*

Abstract | Many organisms that cause infectious diseases, particularly RNA viruses, mutate so rapidly that their evolutionary and ecological behaviours are inextricably linked. Consequently, aspects of the transmission and epidemiology of these pathogens are imprinted on the genetic diversity of their genomes. Large-scale empirical analyses of the evolutionary dynamics of important pathogens are now feasible owing to the increasing availability of pathogen sequence data and the development of new computational and statistical methods of analysis. In this Review, we outline the questions that can be answered using viral evolutionary analysis across a wide range of biological scales.

Balancing selection
Any form of natural selection that results in the maintenance of genetic polymorphisms in a population, as opposed to their loss through fixation or elimination.

**Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.*
‡Institute for Evolutionary Biology, University of Edinburgh, Kings Buildings, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK.
e-mails:
oliver.pybus@zoo.ox.ac.uk;
a.rambaut@ed.ac.uk
doi:10.1038/nrg2583
Published online 30 June 2009

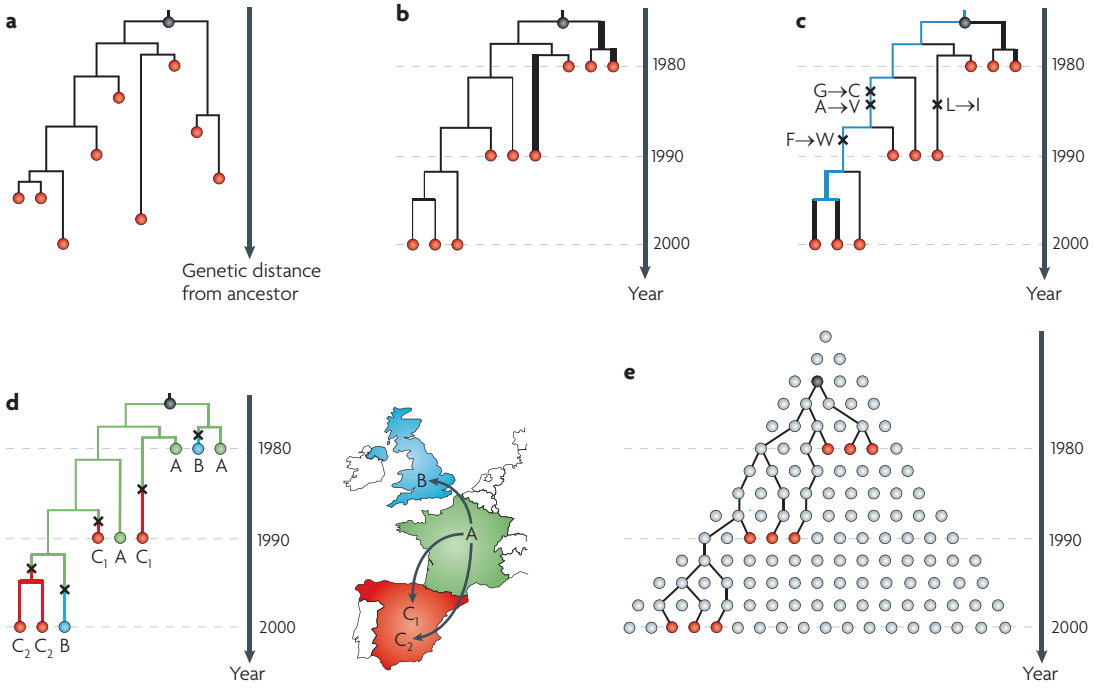
Rapidly evolving pathogens are unique in that their ecological and evolutionary dynamics occur on the same timescale and can therefore potentially interact. For example, the exceptionally high nucleotide mutation rate of a typical RNA virus[1] — a million times greater than that of vertebrates — allows these viruses to generate mutations and adaptations *de novo* during environmental change, whereas other organisms must rely on pre-existing variation maintained by population structure or balancing selection. In addition, many viruses frequently recombine, further increasing the opportunity for genetic novelty. Consequently, populations of fast-evolving pathogens can accumulate detectable genetic differences in just a few days and can adapt brutally swiftly, even when the adapted genotype would have been strongly deleterious in a previous environment. The interaction between evolution and epidemiology is reciprocal: the maintenance of onward transmission may be crucially dependent on continuous viral adaptation, just as the fate of a viral mutant may be decided by its hosts' position in a transmission network.

The term phylodynamics has been coined[2] to describe infectious disease behaviour that arises from a combination of evolutionary and ecological processes, and we adopt the term in this Review as a convenient shorthand for the existence and investigation of such behaviour. We focus on studies that infer viral transmission dynamics from genetic data; these are typically based on concepts from phylogenetics and population genetics, but they also link pathogen evolution to the dynamics of infection and transmission. In the last decade, such studies have matured from theoretical and qualitative investigations (for example, REFS 3,4) to global genomic investigations of key human pathogens (for example, REFS 5–7). Understandably, most studies have focused on important human RNA viruses such as influenza virus, HIV, dengue virus and hepatitis C virus (HCV); therefore, this Review concentrates on these infections. However, the range of pathogens and hosts to which phylodynamic methods are applied is expanding, and we also discuss infectious diseases of wildlife, crops and livestock.

The field of viral evolutionary analysis has greatly benefited from three developments: the increasing availability and quality of viral genome sequences; the growth in computer processing power; and the development of sophisticated statistical methods. Although the explosion in viral genomic data is outpacing our ability to develop methods that fully exploit the potential of these data, we provide an overview of the key biological questions that can be tackled using current evolutionary analysis methods (BOX 1). For example, when did a newly emergent epidemic begin, and from which population or reservoir species did it originate? Can genetic data resolve the order and timing of transmission events during an outbreak? How swiftly do pathogen strains move between continents, regions and epidemiological risk groups, or even between different tissues in a single infected host? Perhaps the most recognizable achievements of viral evolutionary analysis to date are the reconstruction of the origin and worldwide dissemination of HIV-1 (REFS 8,9–16), and the explanation of influenza A epidemics through the combined effects of natural selection and global migration[5,6,17–24].

We describe the range of empirical questions that phylodynamic studies can address by outlining the findings of important studies, most of which have

Box 1 | **Phylodynamic techniques**



Rooted molecular phylogenies can be estimated from viral gene sequences (see the figure, part **a**). Depending on the scale of the analysis undertaken, the sampled sequences (red circles) may represent infected individuals, infected cells, virions or higher-level units such as villages. The phylogeny branching order shows the shared ancestry of the sequences, which usually — but not always — reflects the history of pathogen transmission between these units (discussed in main text). This phylogeny has no timescale, so the branch lengths represent the genetic divergence from the ancestor (black circle). If the sequences of interest undergo recombination, then a single phylogenetic tree may not adequately describe evolutionary history and alternative methods can be applied (for example, REF. 104).

The same phylogeny can also be reconstructed using a molecular clock model (see the figure, part **b**), which defines a relationship between genetic distance and time. The pathogen sequences have been sampled at known time points and the phylogeny branches have lengths in units of years. This approach estimates the ages of branching events, including that of the common ancestor. The simplest, 'strict' clock model assumes that all lineages evolve at the same rate. More complex, 'relaxed' models allow evolutionary rates to vary through time or among lineages, resulting in variation around an average rate[25]. In this phylogeny, unusually fast or slow evolving lineages are shown as thick or thin lines, respectively. The relationships among genetic distance, evolutionary rate and time can be understood by comparing the branch lengths in part **a** and part **b**.

Phylodynamic data can also highlight the evolution through time of mutations that may reflect viral adaptations (see the figure, part **c**). Observed amino acid changes (crosses) are shown mapped onto specific phylogeny branches. Amino acid sites under positive selection can be identified using dn/ds methods, which compare the rate of replacement substitutions (that change the amino acid) with the rate of silent substitutions (that do not change the amino acid)[18,105]. Such methods are most powerful when detecting diversifying selection, making them appropriate for the analysis of infectious disease, but the results obtained using these methods require careful interpretation[106]. Of particular interest are the replacement mutations that are found on the persisting phylogenetic 'backbone' that represents the ancestor of future virus populations (blue branches), as opposed to those occurring on branches that die out (black branches).

The data can also be analysed using temporal phylogeography (see the figure, part **d**). The nine sequences were sampled from France (green, A), the United Kingdom (blue, B) and two locations in Spain (red, C_1 and C_2). Statistical methods can be used to reconstruct the history of pathogen spread, so that each branch is labelled with its estimated geographic position. Current reconstruction methods mostly use simple parsimony approaches[107] that reconstruct a minimum set of migration events consistent with the observed phylogeny. Lineage movement events are marked on the phylogeny with crosses. Combining the spatial and temporal information provides further insights — this hypothetical pathogen spread to location C_1 years before independently arriving at location C_2. Such analyses are not limited to hypotheses concerning physical geography, as the labels A, B, C can stand for any trait of interest, for example, host species, cell tropism during infection, host risk factors or clinical outcome.

The principles of coalescent analyses, which incorporate an explicit model of the sampled pathogen population, are illustrated in figure, part **e**. Each circle represents an infection, and circles on the same row occur during the same period of time. The increasing width of each row therefore reflects the growth of the epidemic through time. Starting from the sampled infections (red), the sampled lineages (black lines) can be traced back through unsampled infections (grey) to the common ancestor (black circle). The rate at which the sampled lineages merge or coalesce depends on population processes such as population dynamics, population structure, selection and recombination (only change in population size is represented here). Coalescent methods are used to infer these processes from randomly sampled pathogen sequences.

Diversifying selection
Any form of natural selection that generates high levels of genetic diversity; for example, recurrent positive selection or balancing selection.

Parsimony approach
A principle of evolutionary inference, based on the assumption that the best-supported evolutionary history for a characteristic is the one that requires the fewest number of changes in that characteristic.

been published in the last few years. Our Review also highlights the variety of practical contexts in which such questions arise, including epidemic management and control, understanding variation in clinical disease, the design of effective vaccines, and criminal trials in which negligent transmission has been alleged. To emphasize the general applicability of the phylodynamic approach, we consider the various organizational scales at which analyses are undertaken, from the global evolutionary behaviour of pathogens to evolution in a single infected host. It is clear that, even for the same pathogen, evolutionary and ecological processes combine in different ways at different scales[2] (BOX 2). For example, influenza A virus displays strong genetic evidence of antigenic selection when studied over many years, but seems to be dominated by stochastic processes when only a single epidemic in one location is considered[22]. We also discuss aspects of data collection, pathogen biology and analysis methodology that may promote or hinder the generation of reliable conclusions.

Methods to analyse viral evolutionary dynamics

Investigating the joint evolutionary and ecological dynamics of infectious disease requires a common frame of reference within which models and data from different fields can be integrated. As we illustrate, this is often achieved by reconstructing evolutionary change on a natural timescale of months or years, enabling researchers to date epidemiologically important events such as zoonotic transmissions. A real timescale also allows pathogen evolution to be directly compared with known surveillance or time series data, perhaps revealing the time period during which a pathogen existed in a population before its discovery, or indicating the impact of public health interventions on viral genetic diversity.

Phylodynamic analyses commonly use molecular clock models to represent the relationship between genetic distance and time (BOX 1). Early simplistic models that assume a constant rate of virus evolution have been superseded by those that explicitly incorporate rate variation, either between strains or through time (for example, REF. 25).

A second, and increasingly popular, common frame of reference is provided by the geographic or spatial distribution of disease isolates (BOX 1). Combined spatial and genetic analyses not only reveal the location of origin of emerging infections, but can also discern the route of transmission and the rate of geographic spread. In addition, statistical models based on coalescent theory are used to directly link patterns of genetic diversity to ecological processes, such as changing population size and population structure (BOX 1). Using these models, it becomes possible to infer the characteristics of pathogen populations, such as their rate of growth, from a small sample of genomes. The resolution and scope of phylodynamic methods depends on the rate of pathogen evolution relative to that of ecological or spatial change — epidemics that fluctuate faster than mutations accumulate among pathogens will not leave an imprint in genetic diversity, although longer-term dynamic trends will.

Molecular clock
A statistical model that describes the relationship between time and the genetic distances among nucleotide sequences. In contrast to older molecular clock models, contemporary models no longer require the assumption that the rates of nucleotide change are constant through time.

Coalescent theory
A theory that describes the shape and size of genealogies that represent the shared ancestry of sampled genes. It describes how the statistical distribution of branch lengths in genealogies depends on population processes such as size change and structure.

Reassortment
A form of genome recombination occasionally exhibited by viruses, such as influenza, which have a genome composed of multiple RNA molecules (genomic segments). The resulting virus produced by reassortment possesses a mixture of genomic segments from two or more parental viruses.

Selective sweep
The rapid increase in frequency of a mutation owing to positive selection for that mutation.

Dynamics on a global scale

The broadest perspective on the evolutionary dynamics of a pathogen is obtained by sampling its worldwide genetic diversity over a suitable period of time. Not all viruses are geographically widespread — some might be limited by the range and dispersal of their hosts — but for those that are, it is essential to understand the geographic structure of viral genetic diversity. For example, HCV shows genotype-specific responses to antiviral drugs, and the clinical severity of dengue virus infection may depend on previous exposure to genetically distinct strains. Genetic data also reveal the rate and route of global spread, which have been most effectively studied for highly infectious airborne viruses such as severe acute respiratory syndrome (SARS) coronavirus and influenza viruses.

Influence of human movement. Humans are an atypical host species as urban population densities and international transport provide opportunities for pathogen transmission that would be otherwise absent. The role of contemporary human migration in determining global viral dynamics has been most comprehensively studied for the influenza A virus by the systematic collection, sequencing and analysis of thousands of viral isolates. Historically, influenza has caused intense bursts of human mortality, most notably associated with the reassortment of human and non-human influenza viruses, which creates strains for which humans have no acquired immunity. Evolutionary analysis of the antigenic haemagglutinin gene (*HA*) of the dominant H3N2 strain has shown that the influenza A virus evolves rapidly through time, yet viruses sampled concurrently from different continents exhibit limited diversity and are typically descended from a common ancestor only a few years earlier[5,22]. Recent evolutionary studies have revealed that the virus re-emerges each year from a persistent Southeast Asian 'source' and follows global aviation networks to temperate 'sink' regions, seeding new winter epidemics there that die out over summer[5,6] (FIG. 1). The global restriction on the diversity of influenza A virus is caused by selective sweeps driven by the host's acquired immunity, which generates rapid antigenic evolution[24] and corresponding high rates of amino acid change at HA antigenic sites[19]. Evolution of influenza A virus is even more dynamically complex when the whole genome is considered — reassortment between genome segments modulates the action of selection, so that some selective sweeps are genome-wide, whereas others only restrict the diversity of *HA*[5].

Reconstructing histories of epidemics. Influenza A dynamics are clearly the result of intricate and ongoing interactions between evolutionary and ecological processes. However, not all pathogens with a worldwide distribution show such complex behaviour at this scale. Although the HIV-1 pandemic is truly international, it is the result of simpler ecological processes that are less strongly coupled to viral evolution. Evolutionary analysis has proven successful in reconstructing the global epidemic history of HIV-1. Viral sequences sampled at

Box 2 | **Linking evolutionary scales: HIV as an example**

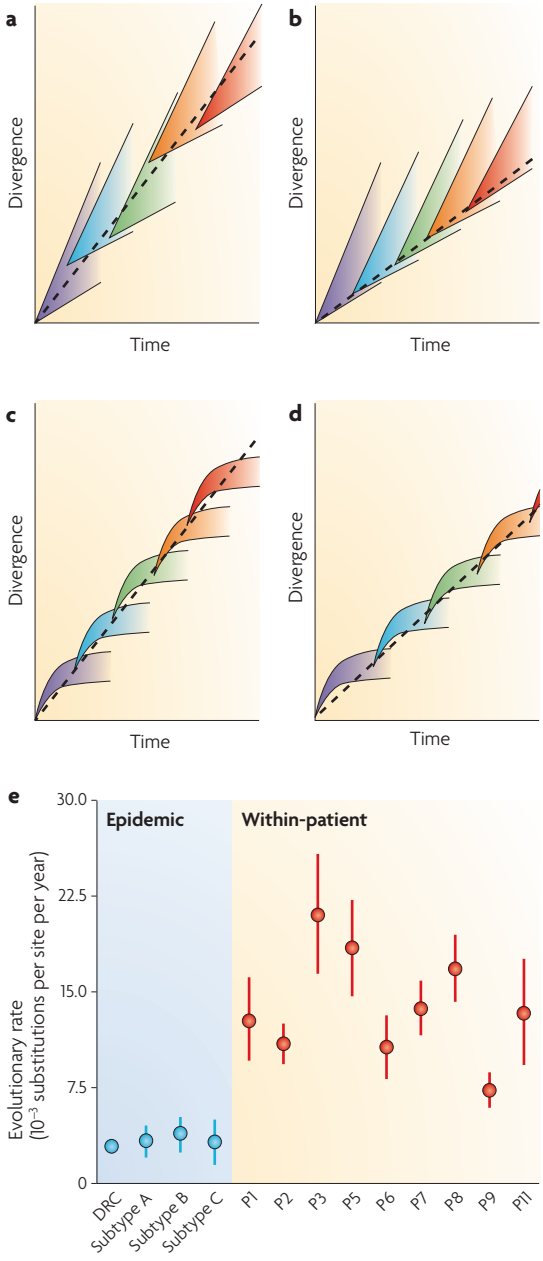
To illustrate the challenges involved in understanding dynamics at multiple levels jointly, we consider here the well-characterized rate of HIV-1 genome evolution at the within-host and between-host scales. The divergence rates of a series of infections can be plotted against time (see the figure, parts **a**–**d**). Each infection is represented by a differently coloured cone of divergence — the gradient of each cone equals the mean rate of within-host virus evolution and the width of each cone represents the variance of this rate. The long-term accumulation of virus divergence at the epidemic level (dashed lines) depends on three factors: the variation in evolutionary rate among strains within a host; whether the average viral rate varies over the course of infection; and whether the strain transmitted to the next host is selected randomly with respect to its evolutionary rate. Empirical analyses indicate a high variance in evolutionary rate among lineages within a host[75], which is caused, at least in part, by latent non-replicative infection of cells[108]. Provided that the lineages are transmitted to subsequent hosts randomly (see the figure, part **a**), the long-term virus evolutionary rate will, on average, equal the average within-host evolutionary rate, even when these average rates differ between patients (P) (see the figure, part **e**).

Discrepancy between within- and between-host rates

In contrast to the above, it seems that HIV-1 evolutionary rates are slower when measured at the epidemic level (see the figure, part **e**; DRC, Democratic Republic of Congo) than when measured at the within-host level[109] (see the figure, part **e**; P1–P9 and P11). One explanation for this difference is that transmission is nonrandom, such that slower-evolving lineages are more likely to successfully generate the next infection than faster ones, with the result that the long-term rate is less than the average within-host rate (see the figure, part **b**). Indeed, the short-sighted action of natural selection will tend to favour those strains with higher within-host fitness, even at the cost of lowered transmissibility. Thus, transmitted viruses could be preferentially drawn from lineages that have accumulated fewer mutations, such as those that have spent a greater proportion of time in a latent state. This effect may be enhanced by the existence of a genetically distinct HIV subpopulation in genital mucosa[88,89].

The discrepancy between within- and between-host rates can also be explained if viral evolutionary rates decrease over the course of infection (see the figure, parts **c**,**d**). Several processes could cause such a decrease: the rate of viral replication declines as the disease progresses[75,110]; selection for viral immune escape variants weakens later in infection[76,105]; and adaptation of the viral population is fastest early in infection, soon after its transmission to a new host environment. As yet, the possible effect of recombination on HIV evolutionary rates at different scales is unknown.

Whatever the underlying cause, if average evolutionary rates vary during infection then the long-term rate of evolution becomes dependent on when transmission occurs. If within-host rates decline during infection then more rapid transmission will result in a faster long-term rate of evolution (see the figure, part **c**) than slower transmission (see the figure, part **d**). This has been shown for the human T cell lymphotropic virus type II, a leukaemia-causing relative of HIV, which seems to evolve many times faster in rapidly transmitting drug users than in populations that are vertically infected during breastfeeding[4]. Conversely, it has been argued that within-host rates increase over the first weeks of infection, owing to the activation of the immune response that drives viral adaptation, hence fast early transmission could alternatively lead to slower long-term rates[111].



various times since the discovery of HIV in 1983 have been used to date the origin of the pandemic to the first half of the twentieth century[10,15] and to pinpoint west-central Africa as its geographic source[14]. These results have been validated and refined by the recovery of genomic fragments from older isolates, notably two 50-year old preserved tissue samples from Kinshasa, Democratic Republic of Congo[15,16], which indicate that considerable HIV diversity had accrued there by 1960.

The worldwide dissemination of HIV-1 from its central African source over several decades was propelled by multiple 'founder events', whereby individual HIV-1 lineages moved to new regions and established epidemics, sometimes recombining in the process, thus generating an array of circulating recombinant forms. The nature and timing of both founder and recombination events have been estimated by evolutionary analysis[8,13,26]. In contrast to influenza A, the absence of protective immunity against HIV means that viral adaptation probably played little part in shaping the current geographical distribution of HIV-1 subtypes, although there is evidence that the virus acquired specific mutations after zoonosis to enable efficient transmission among humans[27] and that HIV-1 is now adapting to the diversity of human leukocyte antigen class I molecules[28,29].

Simple epidemic dynamics also explain the global dissemination of HCV, which has infected humans for at least several centuries[30]. A handful of endemic HCV

Zoonosis
An infectious disease transmitted from animals to humans, or the event of cross-species transmission.

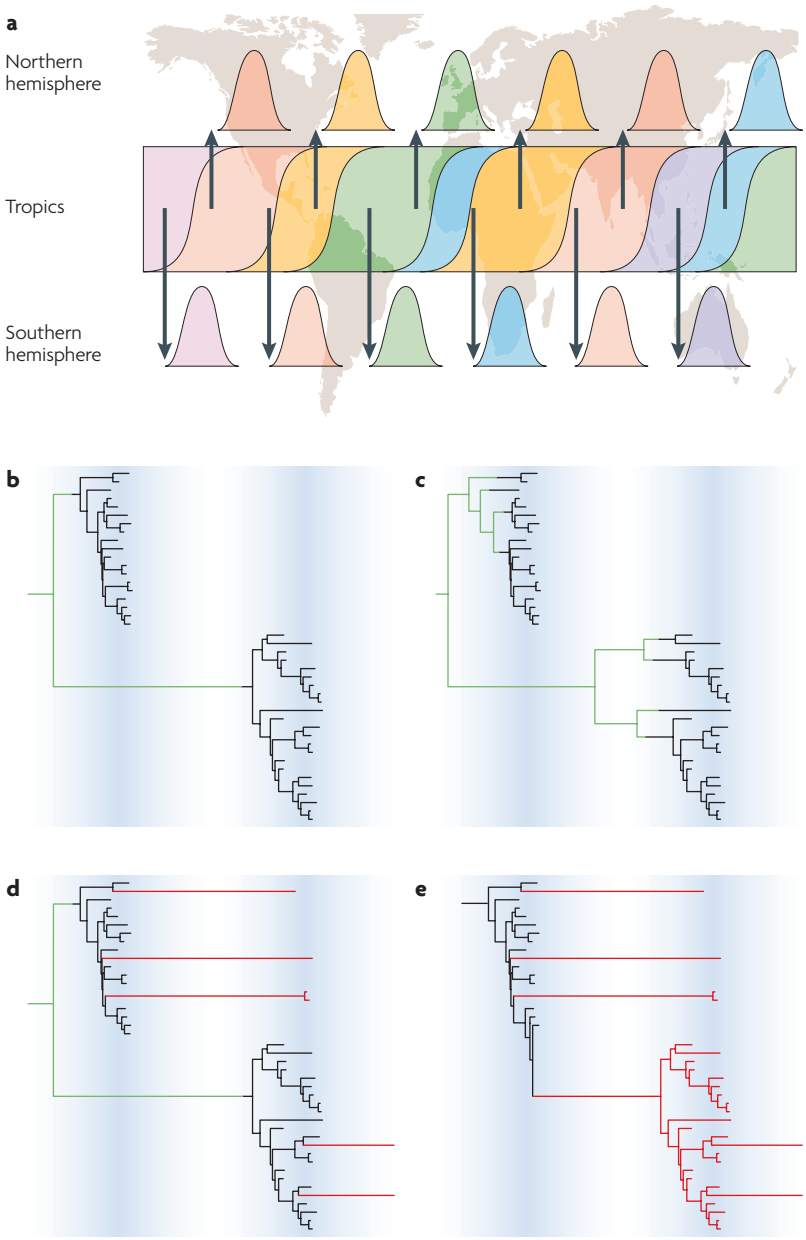


Figure 1 | **The global dynamics of influenza A virus.** Human influenza A virus exhibits a complex pattern of global seasonal dynamics, with epidemics in temperate areas occurring during the winter and year-round sporadic outbreaks in the tropics. Recent analyses indicate that these dynamics are best described by a source–sink model of viral population structure, with a persistent reservoir in South-East Asia driving viral diversity worldwide[5,6]. **a** | Complete genome sequences sampled from New York State, USA, and from Australia and New Zealand have provided a high-resolution snapshot of diversity in these locales over successive seasons[5,22]. Continuous transmission of influenza in the reservoir populations allows natural selection for antigenic diversity, whereas the sink populations with seasonal dynamics will tend to be a representative sample of this diversity. **b**–**d** | Different patterns of global gene flow will be reflected in the phylogenies of influenza isolates sampled from sequential epidemics in one location. **b** | The entire diversity of the second season is descended from a single lineage originating from the global reservoir (lineages representing this global reservoir are in green). **c** | As part **b**, but with multiple lineages from the global reservoir seeding each season. **d** | As part **b**, but with a few lineages persisting locally (red) from one season to the next. **e** | The entire second season is descended from local lineages, implying that transmission persists from season to season in this location. Part **a** is modified, with permission, from REF. 5 © *Nature* (2008) Macmillan Publishers Ltd, all rights reserved.

strains, originally from Asia and Africa, exploded in prevalence worldwide during the twentieth century owing to their chance association with new routes of transmission, such as transfused blood[31].

Emerging insights. Although few pathogens have been sampled as comprehensively as influenza A virus or HIV-1, new insights are being gained as large data sets are compiled for other viruses. For example, recent studies of echovirus 30, a transmissible human enterovirus that causes periodic outbreaks of meningitis, have revealed a fascinating picture of evolutionary forces that vary among viral genes[32,33]. Echoviral capsid genes diverge continuously and rapidly, show rapid global transmission, but exhibit limited concurrent variation. This is analogous to the immune-driven turnover of influenza A *HA* lineages, but there is substantially less genetic evidence of positive selection for immunologically novel echoviral variants[32,33]. By contrast, echovirus 30 polymerase gene lineages are geographically structured, diverse, and coexist on a global scale. Frequent recombination between the capsid and polymerase genes generates transient recombinant forms that are estimated to persist for approximately 5 years[33]. This modular nature of echovirus 30 evolution is all the more remarkable given that it takes place in an unsegmented linear genome that is less than 8 kb long.

Human metapneumovirus, a recently discovered and common cause of childhood respiratory illness, exhibits complex behaviour that is less fully understood. The virus forms several lineages, each of which contains little genetic diversity — suggesting that genetic bottlenecks are common, but only partial or local in effect[34].

Evolutionary analysis has helped track the global spread of the H5N1 highly pathogenic avian influenza (HPAI). Because the virus has been continuously sampled since its emergence in China in 1996, phylogenies can provide accurate reconstructions of its movements, both internationally[35] and locally[36]. Molecular clock results indicate that HPAI lineages typically reside at a location for several months before their official detection[37]. HPAI strains in Asia also undergo frequent reassortment, which may be facilitated by the dense and interconnected duck and poultry populations in the region[36,37].

As more pathogens are studied on a global scale, we should remember that conclusions drawn from small and local samples will underestimate dynamic complexity. Indeed, our understanding of both HIV-1 and influenza virus population dynamics changed appreciably after comprehensive surveys of viral diversity were published[6,14]. If we extrapolate from the examples of echovirus 30 and influenza A virus, then it seems that the most complex global behaviour occurs in highly transmissible viruses that cause acute infections and short-lived epidemics, possibly because their dynamics arise from a three-way interplay between transmission, host herd immunity and viral adaptation. Human viruses that might show such behaviour — when sampled on a sufficiently large scale — include enteroviruses, rhinoviruses, caliciviruses and paramyxoviruses.

Regionally or genetically defined epidemics

A large proportion of evolutionary analyses of pathogens consider individual lineages, strains or subtypes circulating in a specific location, which may be a whole continent or just one town or district. Such outbreaks frequently correspond to a single epidemic, as defined by surveillance organizations, and may involve a single lineage or cluster of infections, as defined by phylogenetic analysis. Evolutionary analysis at this scale can determine the source and time of origin of an epidemic, reveal its genetic composition, and is often used to estimate the rate of viral transmission and spatial spread in the affected region.

Locating the source of an epidemic. Studies on a regionally or genetically defined scale often begin by seeking the source of the new strain, which could be either a zoonotic reservoir or an epidemiologically distinct or distant human population. The origin of an epidemic is typically inferred by finding the most genetically similar non-epidemic strain. This is a simple procedure but is greatly dependent on previous sampling. For example, the SARS coronavirus was highly distinct with no close relatives when initially characterized in April 2003 (REF. 38). The discovery in October 2003 of related viruses in civet cats from animal markets[39] suggested that SARS originated from a zoonotic source, but further sampling has shown that bats are the primary reservoir of these viruses[40]. Molecular clock analysis of bat coronaviruses indicates that the cross-species transfer to civet cats occurred only 4 years before the onset of the human epidemic[41].

Epidemic origins are also hard to locate if the source is geographically or temporally remote; West Nile virus strains sampled from the Mediterranean in 1998 were quickly identified as the source of the 1999 North American epidemic[42], whereas the discovery of the probable zoonotic source of pandemic HIV-1 — *Pan troglodytes troglodytes* chimpanzees in south-eastern Cameroon — was the culmination of many years of research[9].

In some instances, genetic analysis can reveal hidden multiple origins for epidemics that initially seemed homogenous. The 1980s HIV epidemic in the UK among men who have sex with men and the 1990s outbreak of HCV in a subset of the same population are both comprised of at least five distinct strains, each with similar epidemiological behaviours[43,44]. Similarly, phylodynamic analysis of whole viral genomes indicates that the 2005 Singapore dengue virus epidemic comprised multiple viral lineages of different geographical origins[45]. The existence of hidden genetic heterogeneity within an epidemic implies that rapid movement of lineages at a higher geographic scale is likely.

Spatial dynamics. Viral isolates sampled from regional epidemics can contain valuable information about the spatial dynamics of infection. For example, Biek *et al.*[46] estimated the spread of raccoon rabies across the north-eastern United States from sequences sampled over three decades. Viral movement was initially rapid but slowed considerably after a few years as individual lineages became established in different locales, and ecological data on outbreak size closely matched the estimates obtained using coalescent methods (see next section). A similar process of invasion and establishment was also reported for dengue virus in the Americas[47]. Interestingly, dengue virus diversity was maintained across epidemic cycles by the metapopulation structure built up during the invasion phase (FIG. 2). If both the location and sampling date of viral sequences is specified it is possible to estimate the distance pathogens move per year solely from genetic data, as demonstrated by reconstructions of Ebola virus spread in central Africa[48] and feline immunodeficiency virus infection of Rocky Mountain cougars[49].

Coalescent theory analysis. Regionally or genetically defined outbreaks most closely represent the typical 'epidemic' that is described by models of mathematical epidemiology. In some cases this representation can be formalized using population genetic models based on coalescent theory, which directly link phylogenetic structure with ecological processes (BOX 1). This approach is typically used to infer past rates of epidemic growth from sampled viral sequences[3] but can, in some circumstances, be used to directly estimate the fundamental epidemiological parameter (R_0) from such data[30,46,50]. Coalescent-based methods have been successfully applied to HCV and HIV-1. This success is partly because of the chronic nature of infection and the absence of cross-immunity for these viruses, which result in comparatively slow changes in prevalence that leave clear footprints in the patterns of viral diversity. Analysis of HCV genomes indicates that, during the twentieth century, strains varied significantly in their rates of growth according to the transmission route by which each strain was spread[30,51]. The reliability of coalescent-based methods — which make a number of limiting assumptions — was tested in an analysis of HCV in Egypt: here, the methods correctly reconstruct a mid-twentieth century explosion in transmission that was caused by widespread unsafe injection during campaigns against schistosomiasis[52]. Comparable phylodynamic studies of HIV-1 subtypes also show agreement between genetic and epidemiological reconstructions[53,54], even though commonly used coalescent methods ignore the presence of HIV recombination.

As well as describing the origin and spread of many individual outbreaks, analyses of regional epidemics have helped reveal conceptual connections between the different fields of epidemiology, population genetics and phylogenetics, and have validated methods of statistical inference. Despite the choice of examples above, analysis at this scale is not limited to human and animal pathogens. For example, Fargette *et al.*[55] linked the timescale of the emergence of rice yellow mottle virus to the nineteenth century expansion of rice culture in Africa, and Almeida *et al.*[56] used similar methods to conclude that the human transport of contaminated plants disseminated banana bunchy top virus among Hawaiian islands after it was introduced to the islands in 1989.

Positive selection
Also known as directional selection. A form of natural selection that results from an increase in the relative frequency of one genetic variant compared with other variants. It often results in the fixation of the selected variant in the population.

Herd immunity
The protection of susceptible members of a population from infection owing to the sufficiently high prevalence of immune individuals.

Metapopulation structure
A metapopulation is composed of multiple subpopulations, among which there is gene flow. Subpopulations also arise and become extinct dynamically through time.

Fundamental epidemiological parameter
(R_0). The basic reproductive number of an infectious disease, from which many epidemiological predictions can be made. It is equal to the number of secondary infections caused by a single infection in a wholly susceptible host population.

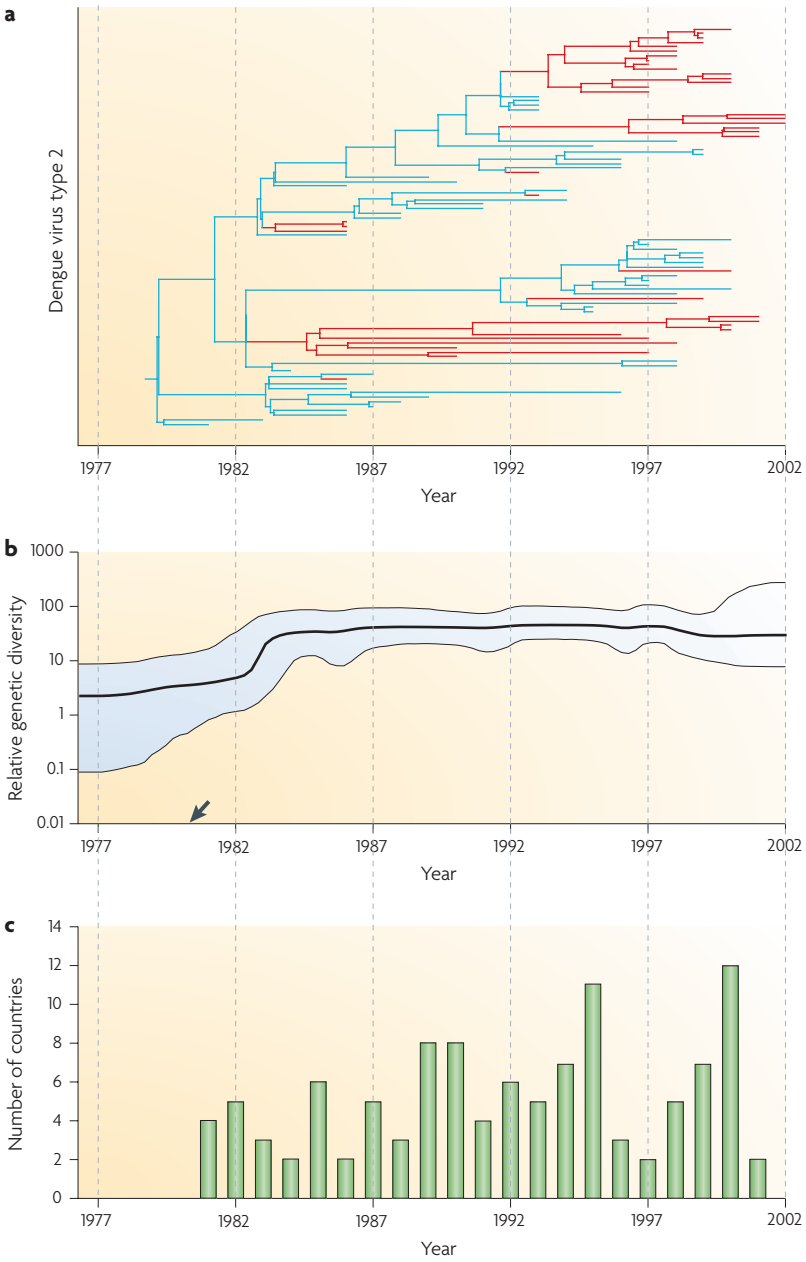


Figure 2 | **A spatially and temporally defined epidemic. a** | A molecular clock phylogeny that illustrates the history of dengue virus genotype 2 infection in the Caribbean and in Central and South America[47]. A simple parsimony approach has been used to reconstruct the likely location of each phylogenetic branch (blue, Caribbean islands; red, mainland Central America and mainland South America). By combining phylogenetic and geographic information, the phylogeny indicates that the outbreak began in the Caribbean before repeatedly and independently invading mainland locations some years later. **b** | An estimate of the relative genetic diversity of the same dengue virus epidemic, which shows an initial increase before stabilizing (95% confidence limits shown in blue). This stabilization does not match the varying number of reported dengue outbreaks (shown in part **c**), probably because spatial population structure maintains viral diversity across epidemic peaks and troughs. More generally, when the sampled population exhibits strong positive selection or population structure then the y-axis cannot be reliably interpreted as proportional to effective population size. The estimated common ancestor of the sampled sequences (arrow) is dated slightly earlier than the first reported outbreak in the region (see part **c**). **c** | Shows the number of countries affected by dengue virus genotype 2 infection per year. Figure is modified, with permission, from REF. 47 © (2005) American Society for Microbiology.

Infection clusters and transmission chains

If an outbreak or infection cluster occurs on a small enough scale then we can realistically expect to sample viruses from all or most of the individuals involved. Studies of such outbreaks tend to fall into two categories: those for which the transmission history (that is, who infected whom, and when) is mostly or wholly known, and those for which it is unknown. Examples in which the transmission history is known are highly informative, as the specified infection history allows evolutionary processes to be investigated with a greater degree of certainty. When the transmission chain is unknown, the primary goal may be the reconstruction of the chain or the identification of its source, timescale or transmission route.

Known transmission histories. Naturally occurring outbreaks for which the transmission event details are known are understandably rare; the majority of those with known details are HIV outbreaks. Known chains of transmission have been used to measure the rate of HIV evolution[57] (BOX 2) and the magnitude of the bottleneck in virus diversity generated at transmission[58]. The Irish anti-D cohort — a well-studied group of HCV-infected women who were accidentally infected with almost identical strains at the same time — has also provided valuable information about variation in viral evolution, host immune selection and disease outcome between patients[59,60]. Using a different HCV transmission cluster, Wrobel *et al.*[61] demonstrated that molecular clock methods can reliably estimate the date that a patient was infected. Transmission chains can also resolve whether the same viral adaptation arises in different hosts (convergent evolution)[62].

Known transmission chains have been used to test whether sequence-based phylogenies match the true history of transmission among epidemiologically connected infections. Although several studies of HIV clusters have reported close agreement[62–64], it is often not appreciated that there are good reasons to expect occasional mismatches between the phylogeny and the true transmission history of a cluster. When one 'donor' infection transmits the virus to multiple recipients, the common ancestors of viral lineages sampled from the recipients will exist in the donor. If the amount of viral diversity in the donor is comparatively high, then the relative order of phylogenetic splitting events (one for each common ancestor) may differ from the order of infection events (FIG. 3). The branching order of transmission for genetically diverse infections is therefore best analysed using metapopulation models that integrate the process of transmission with that of lineage coalescence[65]. This issue is not only restricted to specialized phylogenetic studies — evolutionary analyses of transmission chains are presented in criminal proceedings in which individuals are accused of intentional or negligent transmission[66].

Reconstructing transmission histories. A new and interesting approach to the analysis of transmission chains is presented in recent studies of UK outbreaks of foot and mouth disease virus (FMDV). These studies describe

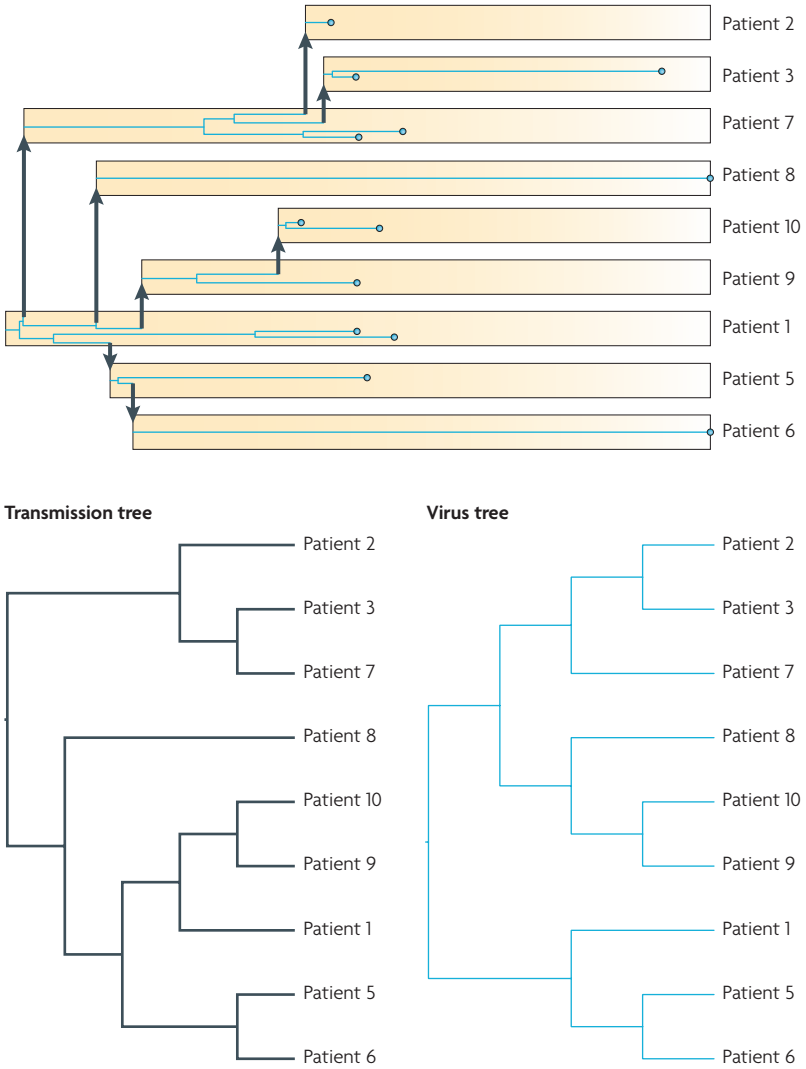


Figure 3 | **Reconstruction of a known HIV-1 transmission chain.** A phylogeny of 13 HIV-1 viral particles (blue circles) sampled at different times (horizontal axis) from 9 different patients for whom the times and direction of viral transmission are known. The virus phylogeny (blue lines) can be mapped within the transmission tree (yellow boxes and arrows), analogous to the mapping of a gene genealogy within a species tree. We can trace all the viruses sampled from one patient back to the time of transmission. Whether more than one lineage is transmitted at this time from the donor will depend on the size of the genetic bottleneck at transmission. Even in the presence of a tight bottleneck, a diverse population in the donor can result in lineage sorting, with the result that the topology of the virus phylogenetic tree does not exactly match the transmission tree.

the infection process at the level of individual farms, with transmission between farms mainly caused by the transport of infected livestock. Cottam *et al.*[67] developed dynamic models that provide a probability distribution for the date of infection of a particular infected farm and its likely period of 'infectiousness' before FMDV diagnosis and culling of the animals. This temporal information was then combined with the genome sequences of viruses that were sampled from the infected herds to identify the most likely chains of transmission linking the farms in time and space. A joint analysis was particularly suitable because FMDV spread is so rapid that comparatively few genetic changes accrue between inter-farm transmissions.

Not all studies of infection clusters focus on the pathways of transmission; sometimes the initiation date of an outbreak is of most interest[68] and at other times the precise epidemic source is sought[69]. However, coalescent-based estimates of population processes are not suitable for infection clusters because this approach requires that the sequences analysed represent a small fraction of the sampled population. Despite this restriction, transmission chain phylogenies can still provide important information about populations, such as the minimum time between transmission events[70]. Furthermore, modern sequencing technology is fast enough for genetic analysis to assist contact tracing and control as an epidemic unfolds. For example, phylogenies confirmed epidemiological suspicions that the 2007 Italian chikungunya outbreak originated from an Indian index case[71]. Considered together, the studies discussed in this section highlight the relevance of transmission chain analyses to applied problems in clinical medicine, forensics and public health. The microevolutionary dynamics of infection events will become a major focus of infectious disease research as high-resolution longitudinal studies will be made possible by the application of next-generation sequencing.

Within-host dynamics

The exceptionally rapid rate of evolution of RNA viruses means that viral evolution in a single host can be studied for the duration of an infection. Dynamics at this scale are fundamental as within-host evolution is the ultimate source of all viral genetic diversity, and therefore it must be understood before models that link different evolutionary scales can be properly developed (BOX 2). Additionally, within-host analyses can reveal the evolutionary processes that underlie some aspects of clinical disease. In practice, such analyses have so far been limited to viruses that establish chronic infections lasting months or years, and for which measurable amounts of genetic change occur between viral samples; this is particularly the case for HIV infection and, to a lesser extent, for HCV and hepatitis B virus[72] infection.

Strong natural selection is clearly the dominant force determining HIV evolutionary dynamics in hosts: HIV phylogenies display a high turnover of short-lived lineages that is driven by host immune selection, analogous to the pattern observed for influenza A virus at the global scale[2] (BOX 2). Correspondingly, HIV genetic diversity at any particular time is low but slowly increases over the course of chronic infection[73]. Numerous analyses have quantified HIV adaptation and evolution using gene sequences, particularly for the viral envelope gene. These studies have found that these processes correlate with the rate of progression to clinical AIDS[74–76] and the rate at which HIV evades neutralizing antibody responses[77]. Equivalent studies of HCV infection have found that viral adaptation predicts the outcome of acute infection[78,79] and that HCV diversity correlates with levels of liver damage[80]. Perhaps the most important outcome of HIV within-host evolution is the generation of T cell escape mutants that can elude host cytotoxic T lymphocyte responses[81] — this is a major barrier to the development

Index case
The first infection in an epidemic or outbreak, from which all subsequent infections are ultimately descended.

of effective HIV vaccines. Although much of the work on T cell escape is not explicitly phylogenetic, there has been a trend away from cross-sectional surveys of viral variation (for example, REF. 82) towards longitudinal and evolutionary studies at all organizational scales, from the level of the pandemic[83] to that of small transmission chains[81] and in individual hosts[84]. The rate at which HIV evolves during an infection depends not only on viral adaptation but also on the replication rate of the virus and its population size: these factors combine to generate measurable variation in viral evolutionary rate both within and between hosts. As a result, evolutionary rates estimated from sequence data may be crucially dependent on the scale of analysis (BOX 2).

Spatial dynamics at the cellular level. Phylodynamic methods have detected and measured the compartmentalization of viral lineages into specific tissues during chronic infection, which creates within-host subpopulations (so-called virodemes), which are analogous to the location-specific clusters of infection seen at higher scales. Highly distinct strains of HIV are found in the brains of patients with neurological illness[85,86], suggesting that virus movement across the blood–brain barrier is not common and might be unidirectional. Finer genetic structure is apparent even among viruses from different brain regions, which seem to evolve at different rates[87]. HIV subpopulations in other tissues have been proposed, including in the cervix[88] and seminal fluid[89], as has compartmentalization in livers with chronic HCV infection[90].

Integrating levels of phylodynamic processes

The evolutionary and ecological dynamics of viral pathogens take place in a hierarchy of organizational scales, from within-host processes to the global dynamics of pandemics, but it is not obvious how dynamics at lower scales combine to generate higher-order behaviour. Such hierarchical processes can be studied from the perspective of both populations genetics[65] and mathematical epidemiology[91]. Multiscale interactions are of great public health importance as well as being of theoretical interest; for example, the success of antiviral drug treatment campaigns will depend on the degree to which drug resistance mutations that arise in treated hosts can accumulate at the epidemic level[92].

There are intriguing parallels between processes in hosts and those at the epidemic or global level[2]. First, within-host studies reconstruct the dynamics of large viral populations from small samples, hence techniques commonly applied to large-scale epidemics (particularly coalescent models) can be re-employed with an appropriate change in perspective — each sequence represents an infected cell or virion, rather than an infected host. Secondly, within-host evolution is closely intertwined with ecological processes, such as the turnover of virions, host cells and components of the host immune response. These dynamics are studied using virus kinetics models[93], which were directly inspired by related models developed by mathematical epidemiologists. As at higher scales, within-host studies have attempted to integrate evolutionary and ecological processes[94–96]; for example, *in vivo* HIV cell-to-cell generation times can be accurately estimated by coalescent analysis of sampled virus sequences[97,98]. There is great potential for further development of models that combine the abundant longitudinal data on infection kinetics with those on viral evolution.

Conclusions

The field of infectious disease evolutionary dynamics is currently seeing a revolution in all three of the technologies on which it relies: genomic sequencing, statistical methodology and high-performance computing. This confluence has produced a burgeoning interest in the evolutionary and epidemiological processes that leave their imprint on pathogen genomes, as reflected in the empirical studies and analysis techniques reviewed here. However, it is our opinion that many investigations still fail to fully appreciate or utilize the rich source of epidemiological information contained in viral genome sequences. Genetic data can independently corroborate surveillance data during an epidemic and can shed light on events before the initial report of the outbreak. Furthermore, evolutionary and surveillance data provide alternative perspectives on the same underlying phylodynamic process and can therefore be validated against one another. The practicality of this approach was demonstrated during the H1N1 'swine flu' epidemic, first detected in April 2009. Tens of viral sequences were made publically available within days of discovery of the virus, and evolutionary analysis was incorporated into initial assessments of the pandemic potential of the new strain[50].

Large-scale sampling and sequencing could also revolutionize our understanding of medically important RNA viruses, such as caliciviruses, rotaviruses and enteroviruses, the genetics of which are currently comparatively neglected. DNA viruses with small genomes that evolve at similar rates to RNA viruses[1] will be equally suitable for phylodynamic analysis. When applied to slower-evolving DNA viruses, bacteria and protozoa, evolutionary analyses similar to those introduced here can help elucidate longer-term processes, such as host–pathogen co-divergence and pathogen speciation[99–101].

In the near future, the greatest impact on viral evolutionary analysis will come from the increasing accessibility of new high-throughput sequencing technologies[102]. For RNA viruses, which have genomes that are on average only 15,000 nucleotides long, it is likely that hundreds or thousands of complete genomes sampled from both viral epidemics and infected hosts can be routinely subjected to molecular epidemiological analysis. Ensuring that computational and statistical developments keep pace with this revolution in data acquisition will be a great challenge. One promising solution is to harness the power of 'multi-core' or massively parallel computing technologies in evolutionary analysis[103]. The coming genomic era will also allow us to determine how much information can be inferred from gene sequences alone — only those ecological processes that occur on the same timescale as genetic change will leave their mark on genetic data, and robust evolutionary inferences carry a statistical uncertainty that should be accurately estimated and reported.

Cross-sectional survey
An investigation that samples a population at a specific point in time. A longitudinal survey, by contrast, samples a population at several different times.

Bayesian inference
A method of statistical inference that uses Bayes' theorem to calculate the probability of a hypothesis. Such methods combine prior information with new observations or data.

Therefore, a clear goal for the future is to further develop analytic methods that combine genetic and epidemiological data to reconstruct epidemic history and to predict future trends, a task to which Bayesian inference methods of statistical inference are well suited. Further development of analysis methods is required in three key areas: the quantification of viral adaptation by natural selection; the explicit integration of evolutionary and spatial information; and the measurement of rates of viral reassortment or recombination. Advances in these areas could raise new questions for phylodynamic analysis. For example, do lineages differ in their rates of spatial diffusion? And are bursts of viral adaptation associated with recombination events? However, such analytical finesse is of little use if basic epidemiological information, such as the date and location of sampling, is unavailable, and we implore researchers generating viral sequences to attach as much sample information to each sequence as ethical constraints permit.

1. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Rev. Genet.* **9**, 267–276 (2008).
2. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
3. Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P. & Harvey, P. H. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos. Trans. R. Soc. Lond. B* **349**, 33–40 (1995).
An early exposition of the idea that pathogen gene sequences contain information about the epidemic history of infectious disease.
4. Salemi, M. *et al.* Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. *Proc. Natl Acad. Sci. USA* **96**, 13253–13258 (1999).
5. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619 (2008).
6. Russell, C. A. *et al.* The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320**, 340–346 (2008).
References 5 and 6 generate new insights into the global evolutionary dynamics of human influenza A virus by analysing thousands of influenza gene sequences collected worldwide.
7. Bird, B. H., Khristova, M. L., Rollin, P. E., Ksiazek, T. G. & Nichol, S. T. Complete genome analysis of 33 ecologically and biologically diverse Rift Valley fever virus strains reveals widespread virus movement and low genetic diversity due to recent common ancestry. *J. Virol.* **81**, 2805–2816 (2007).
8. Gilbert, M. T. P. *et al.* The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl Acad. Sci. USA* **104**, 18566–18570 (2007).
9. Keele, B. F. *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**, 523–526 (2006).
10. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796 (2000).
11. Lemey, P. *et al.* The molecular population genetics of HIV-1 group O. *Genetics* **167**, 1059–1068 (2004).
12. Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl Acad. Sci. USA* **100**, 6588–6592 (2003).
13. Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. & Holmes, E. C. Human immunodeficiency virus — phylogeny and the origin of HIV-1. *Nature* **410**, 1047–1048 (2001).
14. Vidal, N. *et al.* Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* **74**, 10498–10507 (2000).
15. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664 (2008).
16. Zhu, T. F. *et al.* An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597 (1998).
17. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286**, 1921–1925 (1999).
18. Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94**, 7712–7718 (1997).
19. Fitch, W. M., Leiter, J. M., Li, X. Q. & Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proc. Natl Acad. Sci. USA* **88**, 4270–4274 (1991).
20. Holmes, E. C. *et al.* Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.* **3**, e300 (2005).
21. Nelson, M. I. *et al.* Molecular epidemiology of A/H3N2 and A/H1N1 influenza virus during a single epidemic season in the United States. *PLoS Pathog.* **4**, e1000133 (2008).
22. Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A. & Holmes, E. C. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.* **3**, 1220–1228 (2007).
23. Nelson, M. I. *et al.* Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog.* **4**, e1000012 (2008).
24. Smith, D. J. *et al.* Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376 (2004).
25. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PloS Biol.* **4**, 699–710 (2006).
26. Tee, K. K. *et al.* Estimating the date of origin of an HIV-1 circulating recombinant form. *Virology* **387**, 229–234 (2009).
27. Wain, L. V. *et al.* Adaptation of HIV-1 to its human host. *Mol. Biol. Evol.* **24**, 1853–1860 (2007).
28. Kawashima, Y. *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* **458**, 641–645 (2009).
29. Kosakovsky Pond, S. L. *et al.* Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput. Biol.* **2**, e62 (2006).
30. Pybus, O. G. *et al.* The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323–2325 (2001).
Explicitly links epidemiological and population genetic models for the first time, thereby demonstrating that the basic reproductive number of a virus can be estimated from genetic data.
31. Pybus, O. G. *et al.* Genetic history of hepatitis C virus in East Asia. *J. Virol.* **83**, 1071–1082 (2009).
32. Bailly J. L. *et al.* Phylogeography of circulating populations of human echovirus 30 over 50 years: nucleotide polymorphism and signature of purifying selection in the VP1 capsid protein gene. *Infect. Genet. Evol.* **9**, 699–708 (2009).
33. McWilliam Leitch, E. C. *et al.* Transmission networks and population turnover of echovirus 30. *J. Virol.* **83**, 2109–2118 (2009).
34. de Graaf, M., Osterhaus, A. D. M. E., Fouchier, R. A. M. & Holmes, E. C. Evolutionary dynamics of human and avian metapneumoviruses. *J. Gen. Virol.* **89**, 2933–2942 (2008).
35. Wallace, R. G., Hodac, H., Lathrop, R. H. & Fitch, W. M. A statistical phylogeography of influenza A H5N1. *Proc. Natl Acad. Sci. USA* **104**, 4473–4478 (2007).
36. Lam, T. T. *et al.* Evolutionary and transmission dynamics of reassortant H5N1 influenza virus in Indonesia. *PLoS Pathog.* **4**, e1000130 (2008).
37. Vijaykrishna, D. *et al.* Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PloS Pathog.* **4**, e1000161 (2008).
38. Peiris, J. S. M. *et al.* Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
39. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
40. Li, W. D. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
41. Hon, C. C. *et al.* Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J. Virol.* **82**, 1819–1826 (2008).
42. Lanciotti, R. S. *et al.* Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286**, 2333–2337 (1999).
43. Danta, M. *et al.* Recent epidemic of acute hepatitis C virus in HIV-positive men who have sex with men linked to high-risk sexual behaviours. *AIDS* **21**, 983–991 (2007).
44. Hue, S., Pillay, D., Clewley, J. P. & Pybus, O. G. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl Acad. Sci. USA* **102**, 4425–4429 (2005).
45. Schreiber, M. J. *et al.* Genomic epidemiology of a dengue virus epidemic in urban Singapore. *J. Virol.*, **83**, 4163–4173(2009).
46. Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E. & Real, L. A. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc. Natl Acad. Sci. USA* **104**, 7993–7998 (2007).
Shows the application of evolutionary analysis to wildlife disease and provides an excellent example of how epidemiological, spatial and genetic data can be combined.
47. Carrington, C. V. F., Foster, J. E., Pybus, O. G., Bennett, S. N. & Holmes, E. C. Invasion and maintenance of dengue virus type 2 and type 4 in the Americas. *J. Virol.* **79**, 14680–14687 (2005).
48. Walsh, P. D., Biek, R. & Real, L. A. Wave-like spread of Ebola Zaire. *PloS Biol.* **3**, e371 (2005).
49. Biek, R. *et al.* Epidemiology, genetic diversity, and evolution of endemic feline immunodeficiency virus in a population of wild cougars. *J. Virol.* **77**, 9578–9589 (2003).
50. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* 11 May 2009 (doi:10.1126/science.1176062).
51. Tanaka, Y. *et al.* A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *Proc. Natl Acad. Sci. USA* **99**, 15584–15589 (2002).
52. Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. & Rambaut, A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* **20**, 381–387 (2003).
53. Deng, X., Liu, H., Shao, Y., Rayner, S. & Yang, R. The epidemic origin and molecular properties of B′: a founder strain of the HIV-1 transmission in Asia. *AIDS* **22**, 1851–1858 (2008).
54. Paraskevis, D. *et al.* Increasing prevalence of HIV-1 subtype A in Greece: estimating epidemic history and origin. *J. Infect. Dis.* **196**, 1167–1176 (2007).
55. Fargette, D. *et al.* Diversification of rice yellow mottle virus and related viruses spans the history of agriculture from the Neolithic to the present. *PloS Pathog.* **4**, e1000125 (2008).
56. Almeida, R. P., Bennett, G. M., Anhalt, M. D., Tsai, C. W. & O'Grady, P. Spread of an introduced vector-borne banana virus in Hawaii. *Mol. Ecol.* **18**, 136–146 (2009).
57. Leitner, T. & Albert, J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl Acad. Sci. USA* **96**, 10752–10757 (1999).
58. Edwards, C. T. T. *et al.* Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol. Biol.* **6**, 28 (2006).
59. McAllister, J. *et al.* Long-term evolution of the hypervariable region of hepatitis C virus in a common-source-infected cohort. *J. Virol.* **72**, 4893–4905 (1998).

60. Kenny-Walsh, E. Clinical outcomes after hepatitis C infection from contaminated anti-D immune globulin. Irish Hepatology Research Group. *N. Engl. J. Med.* **340**, 1228–1233 (1999).
61. Wrobel, B. *et al.* Analysis of the overdispersed clock in the short-term evolution of hepatitis C virus: using the E1/E2 gene sequences to infer infection dates in a single source outbreak. *Mol. Biol. Evol.* **23**, 1242–1253 (2006).
62. Lemey, P. *et al.* Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J. Virol.* **79**, 11981–11989 (2005).
63. Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* **93**, 10864–10869 (1996).
64. Paraskevis, D. *et al.* Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. *J. Mol. Evol.* **59**, 709–717 (2004).
65. Wilson, D. J., Falush, D. & McVean, G. Germs, genomes and genealogies. *Trends Ecol. Evol.* **20**, 39–45 (2005).
66. Pillay, D., Rambaut, A., Geretti, A. M. & Brown, A. J. L. HIV phylogenetics — criminal convictions relying solely on this to establish transmission are unsafe. *BMJ* **335**, 460–461 (2007).
67. Cottam, E. M. *et al.* Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. Lond. B* **275**, 887–895 (2008).
An innovative statistical analysis of the UK FMDV outbreak that directly combines genetic sequence data with epidemiological surveillance data.
68. de Oliveira, T. *et al.* Molecular epidemiology — HIV-1 and HCV sequences from Libyan outbreak. *Nature* **444**, 836–837 (2006).
69. Guan, Y. *et al.* Molecular epidemiology of the novel coronavirus that causes severe acute respiratory syndrome. *Lancet* **363**, 99–104 (2004).
70. Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A. & Brown, A. J. L. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PloS Med.* **5**, 392–402 (2008).
71. Rezza, G. *et al.* Infection with chikungunya virus in Italy: an outbreak in a temperate region. *Lancet* **370**, 1840–1846 (2007).
72. Lim, S. G. *et al.* Viral quasi-species evolution during hepatitis Be antigen seroconversion. *Gastroenterology* **133**, 951–958 (2007).
73. Shankarappa, R. *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502 (1999).
Reports a comprehensive set of HIV-1 sequences sampled from nine infected patients, the analysis of which has provided new insights into the evolution of HIV.
74. Wolinsky, S. M. *et al.* Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science* **272**, 537–542 (1996).
75. Lemey, P. *et al.* Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* **3**, 282–292 (2007).
76. Williamson, S. Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.* **20**, 1318–1325 (2003).
77. Frost, S. D. *et al.* Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. *J. Virol.* **79**, 6523–6527 (2005).
78. Farci, P. *et al.* The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* **288**, 339–344 (2000).
79. Sheridan, I., Pybus, O. G., Holmes, E. C. & Klenerman, P. High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J. Virol.* **78**, 3447–3454 (2004).
80. Farci, P. *et al.* Evolution of hepatitis C viral quasispecies and hepatic injury in perinatally infected children followed prospectively. *Proc. Natl Acad. Sci. USA* **103**, 8475–8480 (2006).
81. Leslie, A. J. *et al.* HIV evolution: CTL escape mutation and reversion after transmission. *Nature Med.* **10**, 282–289 (2004).
82. Moore, C. B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296**, 1439–1443 (2002).
83. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**, 1583–1586 (2007).
84. Asquith, B., Edwards, C. T., Lipsitch, M. & McLean, A. R. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells *in vivo*. *PLoS Biol.* **4**, e90 (2006).
85. Wong, J. K. *et al. In vivo* compartmentalization of human immunodeficiency virus: evidence from the examination of pol sequences from autopsy tissues. *J. Virol.* **71**, 2059–2071 (1997).
86. Korber, B. T. *et al.* Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol.* **68**, 7467–7481 (1994).
87. Salemi, M. *et al.* Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of *AIDS*. *J. Virol.* **79**, 11343–11352 (2005).
88. Iversen, A. K. N. *et al.* Preferential detection of HIV subtype C′ over subtype A in cervical cells from a dually infected woman. *AIDS* **19**, 990–993 (2005).
89. Pillai, S. K. *et al.* Semen-specific genetic characteristics of human immunodeficiency virus type 1 *env*. *J. Virol.* **79**, 1734–1742 (2005).
90. Sobesky, R. *et al.* Distinct hepatitis C virus core and F protein quasispecies in tumoral and nontumoral hepatocytes isolated via microdissection. *Hepatology* **46**, 1704–1712 (2007).
91. Mideo, N., Alizon, S. & Day, T. Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends Ecol. Evol.* **23**, 511–517 (2008).
92. Wensing, A. M. *et al.* Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *J. Infect. Dis.* **192**, 958–966 (2005).
93. Nowak, M. A. & May, R. M. *Virus Dynamics* (Oxford Univ. Press, Oxford, 2000).
94. Nickle, D. C. *et al.* Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. *J. Virol.* **77**, 5540–5546 (2003).
95. Kelly, J. K., Williamson, S., Orive, M. E., Smith, M. S. & Holt, R. D. Linking dynamical and population genetic models of persistent viral infection. *Am. Nat.* **162**, 14–28 (2003).
96. Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F. & Hanage, W. P. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 17441–17446 (2007).
97. Rodrigo, A. G. *et al.* Coalescent estimates of HIV-1 generation time *in vivo*. *Proc. Natl Acad. Sci. USA* **96**, 2187–2191 (1999).
A key paper that extended coalescent theory to sequences sampled at different times. It is the first study to apply coalescent theory to virus dynamics at the within-host level.
98. Achaz, G. *et al.* A robust measure of HIV-1 population turnover within chronically infected individuals. *Mol. Biol. Evol.* **21**, 1902–1912 (2004).
99. Falush, D. *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585 (2003).
100. Ehlers, B. *et al.* Novel mammalian herpesviruses and lineages within the Gammaherpesvirinae: cospeciation and interspecies transfer. *J. Virol.* **82**, 3509–3516 (2008).
101. Joy, D. A. *et al.* Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300**, 318–321 (2003).
102. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
103. Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376 (2009).
104. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
105. Bonhoeffer, S., Holmes, E. C. & Nowak, M. A. Causes of HIV diversity. *Nature* **376**, 125 (1995).
106. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
107. Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).
108. Finzi, D. *et al.* Latent infection of $CD4^+$ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Med.* **5**, 512–517 (1999).
109. Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* **8**, 125–140 (2006).
110. Lee, H. Y., Perelson, A. S., Park, S. C. & Leitner, T. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput. Biol.* **4**, e1000240 (2008).
111. Maljkovic Berry, I. *et al.* Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. *J. Virol.* **81**, 10625–10635 (2007).

Acknowledgements
We would like to thank E. Holmes and three referees for commenting on the manuscript and improving it immeasurably. We thank A. Drummond and P. Lemey for providing rates of HIV-1 evolution for FIG. 4. Finally we gratefully acknowledge The Royal Society of London, which supports both authors.

FURTHER INFORMATION
Oliver Pybus' homepage: http://evolve.zoo.ox.ac.uk
Andrew Rambaut's homepage: http://tree.bio.ed.ac.uk
***Nature Reviews Genetics* Series on Modelling:**
http://www.nature.com/nrg/series/modelling/index.html
***Nature Reviews Microbiology* Focus on Influenza:**
http://www.nature.com/nrmicro/focus/influenza/index.html
ALL LINKS ARE ACTIVE IN THE ONLINE PDF