Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Review

CellPress

Omics: Fulfilling the Promise

Supersize me: how whole-genome sequencing and big data are transforming epidemiology

Rowland R. Kao[1], Daniel T. Haydon[1], Samantha J. Lycett[1], and Pablo R. Murcia[2]

[1] Boyd Orr Centre for Population and Ecosystem Health, College of Medical Veterinary and Life Sciences, University of Glasgow, G61 1QH, UK
[2] Medical Research Council (MRC) Centre for Virus Research, College of Medical, Veterinary and Life Sciences, University of Glasgow, G61 1QH, UK

In epidemiology, the identification of 'who infected whom' allows us to quantify key characteristics such as incubation periods, heterogeneity in transmission rates, duration of infectiousness, and the existence of high-risk groups. Although invaluable, the existence of many plausible infection pathways makes this difficult, and epidemiological contact tracing either uncertain, logistically prohibitive, or both. The recent advent of next-generation sequencing technology allows the identification of traceable differences in the pathogen genome that are transforming our ability to understand high-resolution disease transmission, sometimes even down to the host-to-host scale. We review recent examples of the use of pathogen whole-genome sequencing for the purpose of forensic tracing of transmission pathways, focusing on the particular problems where evolutionary dynamics must be supplemented by epidemiological information on the most likely timing of events as well as possible transmission pathways. We also discuss potential pitfalls in the over-interpretation of these data, and highlight the manner in which a confluence of this technology with sophisticated mathematical and statistical approaches has the potential to produce a paradigm shift in our understanding of infectious disease transmission and control.

Contact tracing of infectious pathogens and whole-genome sequencing

Identifying pathways of infectious disease transmission can reveal likely points of control and predict future directions of spread. In combination with mathematical models (see Glossary) they can be used to predict the outcomes of alternative control methods. Central to this is epidemiological tracing to identify 'who infected whom', a crucial component of what is known as forensic epidemiology. Unfortunately, tracing is often made difficult by the effort required and the considerable uncertainties in the possible sources of infection and timings of events. Contact patterns can sometimes be inferred from spatiotemporal proximity, particularly where the host populations are sessile and with short-range contacts {e.g., foot-and-mouth disease (FMD) on farms [1], citrus canker in fruit trees [2], rabies in domestic dogs [3], and hospital infections [4]} or through the identification of relevant risk factors (e.g., needle-sharing or sexual contact for HIV transmission). However, even in these cases the difficulty of identifying the most relevant routes and means of contact limits our ability to characterize the underlying transmission processes.

Antigenic or genetic characterization [e.g., serotyping or multi-locus sequence typing (MLST)] of pathogens is an alternative approach to identifying groups of individuals with closely related infections [5]. Until recently these approaches lacked the resolution for characterizing direct contact. However, high-throughput sequencing (HTS) technology, together with improved ability to extract genetic material more cheaply and from smaller pathogen samples [6,7], now allow mass-scale characterization of virtually entire genomes of whole populations of pathogens (generally referred to as whole-genome sequencing or WGS). This technology typically offers orders of magnitude better resolution compared to earlier typing methods [8]. In addition, the increased availability of dense data characterizing the substrate population (e.g., identification of individuals, social groupings, contacts between groups, spatial organization, species compositions etc., and referred to here as denominator data) [9,10], and the development of powerful computational and analytical tools to organize and interpret large datasets, broadens the potential for application of such data to high-resolution epidemiological problems. Although their usage on a large scale is in its infancy, they share many properties with 'big data' problems in other systems: (i) although highly variable in size, big datasets are typically an order of magnitude or greater larger than what had previously been available, (ii) the proportion and coverage of data on the susceptible population of interest that is captured in the datasets

Corresponding author: Kao, R.R. (rowland.kao@glasgow.ac.uk).
Keywords: Mathematical modeling; Bayesian inference; Pathogen evolution; Forensic epidemiology; Who-infected-whom?.

0966-842X/$ – see front matter
© 2014 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.tim.2014.02.011

CrossMark

Glossary

Clustering: informally, the existence of multiple pathways that can lead to a single destination, and particularly when more than one of the pathways is 'short'. In social network analysis there are several formal definitions, with the most common being related to the simplest possible relationship that fulfils the following concept: the number of triangles in a social network (individuals A, B, and C mutually connected) divided by the number of triples in a social network (A connected to B connected to C, but A need not be connected to C).

Competent host: a species that can be infected by a pathogen and also transmit it.

Denominator data: data that describes the composition of a host population, irrespective of the transmission of an infectious disease. This may include the population number or density, the characteristics of individuals, and the connections between them (e.g., friendship networks or movements of individuals between subpopulations). By contrast, numerator data describe the characteristics of the infected population.

Forensic epidemiology: the science of identifying the characteristics of particular infectious disease outbreaks, in particular as they relate to control and eradication, and for which tracing between individuals is an important component.

High-throughput sequencing: the technological revolution that followed the Sanger sequencing technology that was used to generate the first complete human genome, allowing for mass generation of sequences at increasingly affordable costs. Currently broadly subdivided into next- or second-generation sequencing (Illumina or 454) and now third generation (PacBio).

Horizontal genetic transfer: the transfer of genetic material between organisms in a manner other than traditional reproduction (see also recombination and reassortment).

Maintenance host: a host species in which a pathogen can persist – for practical purposes – indefinitely, including if necessary through the mechanism of a vector species (e.g., mosquitoes for malaria).

Mathematical models: a term for quantitative models of disease transmission using mathematical formulae. Usually implying a mechanistic interpretation, with often non-linear transmission dynamics. There are a wide range of usages within this definition, ranging from the highly restrictive (deterministic models with compact mathematical formulations and preferably analytical solutions) to the catholic (that also incorporate purely individual-based simulations).

Monophyletic: a disease outbreak caused by a single external source. By contrast, a polyphyletic outbreak arises from more than one external source.

Orthogonal processes: two or more processes where the variation in each is statistically independent from the other. For example, beyond their most recent common ancestor, two genealogies are orthogonal provided they do not swap genetic material (e.g., through recombination).

Reassortment: the exchange of genetic information via the transfer of genomic segments, as occurs in influenza. It is a special case of recombination with fixed breakpoints.

Recombination: the exchange of genetic material between two pathogens, resulting in the inclusion of material from one into the other and the production of a 'mosaic' genome.

Relative mutation rate: the mean rate at which mutations accumulate divided by the mean time between consecutive generations of infected individuals. This is an indicator of the likelihood that there will be polymorphisms that are informative for tracing between individuals, but also the likelihood that there will be observable differences between the sampled genealogies and the transmission genealogies.

Reservoir host: a species (usually assumed to be wildlife) that is a maintenance host for a pathogen.

Social network: a form of denominator data, describing a population or populations in terms of the individuals hosts (nodes or equivalently in graph theory, vertices) and the associations between them (links, equivalently edges). Social network analysis includes descriptions of clustering which can introduce ambiguities into tracing.

Spillover host: a species that is neither a maintenance host nor is necessary to maintain the pathogen in combination with other host species.

Synonymous mutation: the replacement of a nucleotide by another that does not cause a change in the amino acid sequence after translation.

Transmission network: a form of numerator data, the complete tree of 'who infected whom' in an outbreak.

Whole-genome sequencing (WGS): the process that uses high-throughput sequencing to describe the entire genome of an organism. Because there are always errors or unknown regions in any genome reconstruction, it is more correctly 'nearly-whole' genome sequencing.

are high, and (iii) the variety of data being captured is extensive. The opportunities presented by big data based on WGS are potentially paradigm-shifting, with existing smaller-scale studies [11,12] hinting at what might be possible with very large datasets. Crucial to this is the integration of non-WGS data into analyses identifying epidemiological pathways because this can lead to a considerable refinement of our understanding of transmission. Although this is often conducted descriptively, 'epidemiological' frameworks are being developed that naturally incorporate genetic data with both denominator data and additional information on the transmission of the pathogen across the affected population. In the remainder of this review we shall consider the role that WGS can play in enhancing our understanding of fine-scale epidemiological contact. We shall highlight the pitfalls that arise if there are multiple likely transmission routes for every true transmission route, and where there are differences between observed phylogenies and transmission networks, including the difficulties of inferring the epidemiological dynamics of multi-host pathogens and emerging infections.

Using WGS for tracing

The majority of mutations for any pathogen will be subject to strong purifying selection, with a small minority being subject to positive selection (and potentially a problematic source of homoplasy). This still leaves substantial numbers of neutral or 'nearly neutral' mutations (i.e., sites subject to only weak selection) [13]. Although such nearly neutral variation may be selected out over longer time scales [14,15], over shorter time scales such as a single epidemic they can be useful markers of pathogen genealogy, provided that phenotypic effects [16] are minimal. These mutations will not necessarily be synonymous because there may be constraints imposed by genetic structure (e.g., RNA secondary structure) and overlapping reading frames (i.e., a synonymous mutation on one frame can be nonsynonymous and selected against in the other) [17]. Polymorphisms in sets of sequences can be compromised by technical issues, including errors in sequencing and bioinformatics, resulting in missed or artefactually added mutations), by reassortment in segmented genomes such as in influenza viruses, and by recombination in non-segmented genomes such as those of retroviruses or bacteria [18,19].

All amplification steps can introduce errors, and the more amplification that is required the more likely that errors will be introduced. The number and nature of the artefacts introduced will therefore depend on the size of the original genetic sample, the laboratory protocols used (including the reagents used to process a given sample), the sequencing technology, and also the analytical tools used, with a lack of agreed quality-control protocols providing an additional layer of uncertainty. The nature of the pathogen itself is also important, with RNA viruses requiring error-prone reverse transcription [20]. Such errors carry identifiable signatures; for example artefacts are more likely to be random and appear at low frequency across replicates, unlike the 'true' mutations because these should almost always appear. Methods to identify and minimize these errors are being identified [21,22].

In the absence of horizontal genetic transfer the genetic distance between sequenced pathogens is usually positively correlated with the number of transmission links between individuals. For tracing contact there would ideally be a unique sequence that is shared by the entire within-host population but, immediately upon transmission, would acquire at least one distinguishing mutation.

Box 1. Phylogenies and transmission trees

At coarse spatial and temporal resolutions the evolutionary relationships between pathogen genes will reflect their epidemiological relationships. Pathogens that are closely related epidemiologically will also be those most closely related to each other genetically. However, it is well known [71] that at finer space–time resolution the particular details of the epidemiological process can begin to decouple the transmission tree from the genealogy. In Figure I, filled circles represent genomes sampled from four particular host individuals (hosts labeled A–D and colored red, blue, black, and green respectively). Unfilled circles represent genomes that were not sampled. Circles immediately adjacent to one another are one mutation different to each other. The color of the line indicates which host the different genomes were in. In **(i)**, the pathogen genome sampled is the genome that was transmitted. We can therefore deduce that the most likely transmission scenario is that A infected B, B infected C, and C infected D. However, if transmission occurred sometime before the time of sampling, such that additional mutations were subsequently incurred, either because the infectious period is long, or the mutation rate very high, then we inevitably become less certain which genome was in which host, and consequently several different transmission trees become consistent with the genetic data. It may be that A infected B, B infected C, and C infected D **(ii)**, but the genetic data are equally consistent with a scenario in which D infected A, B, and C **(iii)**, or indeed several alternative explanations (not shown). It is in these situations that the integration of additional data on the timing and the contact process becomes important for inferring the most likely tree. There are other reasons why the transmission tree and phylogeny may be different – for example there may be insufficient genetic information to distinguish between pathogen from different hosts, or recombination or homoplasy may complicate the relationship between the two.

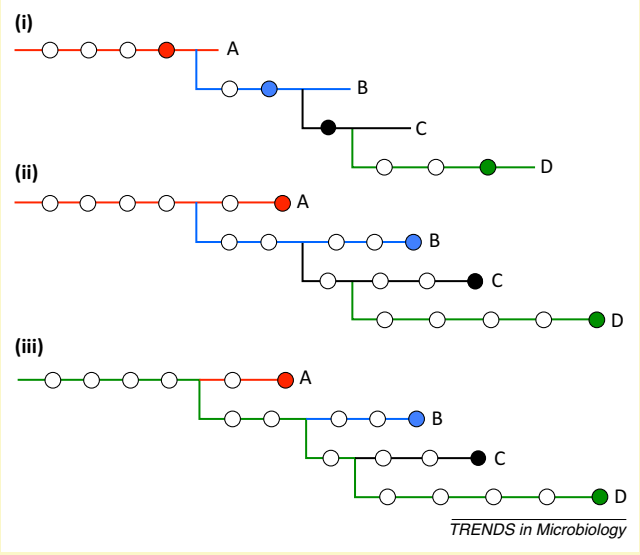

Figure I. A single observed genealogy is consistent with multiple observed transmission processes.

Unfortunately, such a pathogen does not exist, resulting in multiple complications (Box 1). Mutation rates will vary, sometimes considerably [23], as will the times between consecutive transmission events (referred to here as generation times). Further complications arise should epidemiological processes influence evolutionary rates. Examples include the potential role of duration of latency in the mutation rates of *Mycobacterium tuberculosis* in humans [24] and evidence that tropical and subtropical climates accelerate the evolution of bat rabies virus [25]. When mutation rates compared to generation times are low (which we shall call the relative mutation rate), WGS data can provide insufficient genetic signal. For instance, during explosive outbreaks of acute or hyperacute viral infections (i.e., influenza and norovirus), the relative mutation rate might be very low, and consensus sequences among samples from different but closely-related infected individuals may be identical. Such situations may be resolved by using within-host genetic variation to infer properties of between-host transmission [26], as could be achieved by examining the presence of minor allele variants shared by different hosts.

The consideration of within-host dynamics has been shown to reconcile the differences between the sampled phylogeny and transmission dynamics, improving the inferred transmission tree [27,28]. When relative mutation rates are high the differences between time of transmission and sample time, and within-host location of sampling compared to location at which transmission occurs, can increase the observed genetic distances between mother–daughter pairs and introduce ambiguities and biases in the available data (Box 1). For chronic infections such as HIV, hepatitis C, and tuberculosis, an individual could be diagnosed months or even years after transmission and thus the first onward transmission event may pre-date the consensus sequence [29]; this can also be a problem for acute-acting viruses such as FMD virus (FMDV) [30]. The problems are exacerbated where many intermediate cases may be absent from the data, although these may be alleviated by the development of temporal markers of infection or reliable indicators of change in the microbial community [31]. Further difficulties will be incurred if we are interested in transmission processes at multiple scales because the 'ideal' rate would be different at each scale.

The role of model-based inference

Mechanistic models of infectious disease transmission (often called mathematical models) can be used to generate simulated transmission trees based on our understanding of the underlying mechanisms that drive the transmission process (Box 2). Because they are mechanistic, by altering the mechanisms in the model they can be used to predict future outcomes of ongoing epidemics 'how big, and how long?'. It can also be used to predict the outcome of interventions, such as 'will mass vaccination be effective, and what is the required coverage?', 'how many anti-viral drugs will be needed during an influenza pandemic?', or 'is culling or vaccination a better policy to control FMD?'. Such models are designed to provide population-level insights and are typically poor at predicting actual events at small scales because the number of possible transmission trees that result in a given observed epidemic can be very large. When there are even low levels of error and uncertainty in the available data, the accuracy of parameter estimates can be severely degraded [32]. However, good denominator data can substantially reduce the range of possible contacts that result in observed patterns of transmission. Big denominator data are becoming more common; densely sampled associations being recorded include daily, individual movement records for livestock [9,10], and mobile phone network

Box 2. Bayesian model-based inference

Spatial, temporal, and pathogen genetic information have been used in two broadly different ways to reconstruct the dynamics of epidemics. In the first, coalescent models that assume a particular population dynamic model are used to link the demography of the pathogen to its evolution; in this approach a flexible diffusion-like process can be used to estimate the rate of spatial spread of the pathogen [72]. This enables estimation of several useful parameters, including those describing the pathogen demography [73], the diffusivity of the pathogen [74], and the molecular clock [73]. The method is robust to the density of sampling but, because such models are underpinned by fundamentally ecological (or demographic) processes, the estimated parameters do not have straightforward epidemiological interpretations and inferences about high-resolution epidemiological processes are not easily made [75–77]. Indeed, the more highly temporally resolved the data become the more important it is that epidemiological processes are given explicit representation if transmission is to be represented accurately, and the greater the shortcomings that a fundamentally ecological approach has, as opposed to an epidemiological one. Coalescent models can be modified to include an explicit epidemiological focus [41], but they do not as yet account for the high levels of clustering that characterizes spatial spread, for example.

The second approach combines explicit models of transmission with simple models of genetic drift to reconstruct transmission trees reflecting 'who infected whom'. This approach recognizes the host population structure and the epidemiological processes that govern the interaction of host and pathogen. An epidemiological model of disease progression in individuals is used to estimate possible dates of infection and the infectious period of the observed cases. Within this framework the probability of any two cases being causally related can be calculated based on: (i) the probability that the putative donor was infectious and the putative recipient infected during the same time period; (ii) the probability of transmission over the distance separating the two cases; and (iii) the probability that the donor pathogen sequence could have incurred the additional mutations observed in the recipient sequence in the time between collection of the two samples (although the more time that elapses between the transmission of the pathogen and the time that it is sampled from the donor host the more ambiguous inference becomes; see Box 1). This approach allows inferences to be made about latent and infectious periods [12], the transmission tree reflecting 'who infected who' [11,12,49], the rate of evolution 'per transmission event' [30], and the proportion of cases not sampled in a partially observed outbreak [50,77].

[33] and airline traffic data [34,35] for humans. The incorporation of big denominator data into epidemiological models is greatly aided by Bayesian statistical inference frameworks that formalize the relationships between prior knowledge and model-derived likelihood functions. The technical challenges of accomplishing this are not to be underestimated [36], and identifying whether the best-fit model is a good model, particularly where approximate methods have been used, is challenging [37]. These problems are further complicated by the often uncertain relationship between what are often multiple putative routes of transmission and their relative importance to the transmission tree. Although many of the methods used to analyze WGS data are extensions of previous approaches, in one way, WGS provides a unique insight; its unusually dense information can change our understanding of the transmission process at the individual transmission event scale. Also unlike other sources of data, the genealogical relationships are fundamental to the transmission tree, even if the relationship between the transmission trees and the observed genealogies is imperfect (Box 1 and Figure 1). However, ambiguities and errors can be substantially reduced by combining WGS with population-level inference, thereby joining the epidemiological (individual-to-individual) and ecological (demographic, population-level) perspectives [11]. Particular problems are considered below.

The power of the statistical inference will depend on the underlying available data and the robustness of the underlying population growth model [38].The ideal data (i.e., what data points would you ideally have if you were only allowed 'X' samples?) will depend both on the relative mutation rate (e.g., if low, dense sampling of epidemiological clusters may be inefficient) and on the nature of the question being asked. Different questions will require different optimal strategies. For example, the identification of close-scale epidemiological clustering will require extremely dense sampling, which might occur at the expense of coverage of the entire epidemic. This may also include within-host dynamics, for which multiple samples for single individuals may be useful [27,39]. Samples close to the origin of an outbreak will tend to be valuable because they more robustly root the evolutionary analysis. One approach to identifying future samples strategies is to use the inference models themselves to estimate locations and time points where obtaining additional sequences would best improve the inference, or to identify events that appear anomalous under the model and that require further investigation.

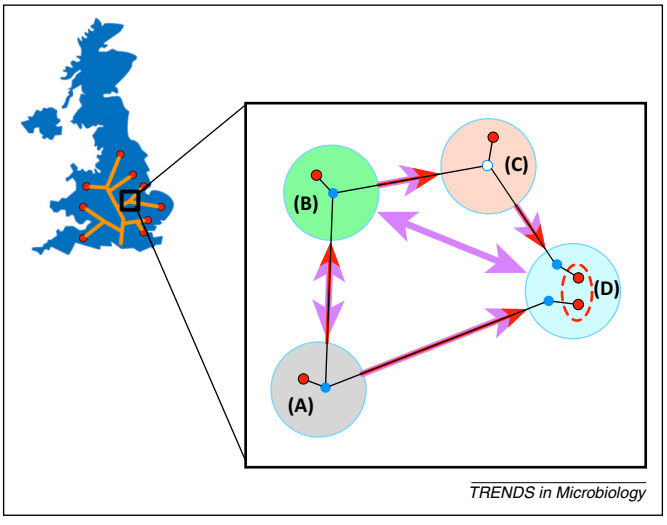

Figure 1. Identifying 'who infected whom' often requires more detailed contact information than is needed at the scales most amenable to phylogeographic approaches (Great Britain scale map, left). At finer scales (right), there are two types of information: genetic information from sampled pathogens (where samples are indicated by red circles) provides direct insight into the transmission network indicated by the red arrows, whereas the possibly bidirectional purple arrows represent the social network (or denominator data). Both help to reconstruct the true transmission tree (red arrows), but deviate from it in different ways. The social network may contain many links that do not cause transmission. By contrast, the transmitted genotypes (blue circles) are indicative of the transmission tree but, especially when mutation rates are low compared to generation times, may lack informative single-nucleotide polymorphisms (SNPs; where the filled circles represent at least one additional mutation, but the open blue circle in C indicates a type identical to what is found in B). A pooled sample from D (broken oval encompassing samples from two lineages) could generate a consensus sequence that is not representative of either transmitting lineage, but these could be recovered by the existence of two divergent sequences that could be identified via deep sequencing.

Clustering, networks, and heterogeneity of potential contacts

Reconstruction of evolutionary pathways largely depends on the identification of common ancestral traits. At a sufficiently coarse geographical scale, infectious disease processes are either spatially correlated or are well described by recordable patterns of interactions such as transportation networks [40]. Although such processes may contain inconsistencies with sampled phylogenetic trees, at these coarse scales they provide the considerable advantage that there is no requirement for either high sampling densities or information about the contact relationships to infer structural properties of the population. Many questions of both scientific and practical interest can be successfully approached this way, provided that there is a sufficiently high density of mutations and possible transmission pathways are largely orthogonal (i.e., few individuals have tightly 'clustered' or shared mutual contacts). By contrast, where transmission occurs over short distances, multiple possible 'non-orthogonal' pathways exist (Figure 1), and thus there can be considerable ambiguity in determining who infected whom. Even at scales where transmission processes are important, if clustering can be discounted, simple compartmental models of disease transmission have been integrated into evolutionary approaches to show in HIV, for example, that phylogenetic clustering (i.e., groups of sequences more closely related than would be expected at random) can be explained by differing phases of transmission intensity [29,41].

At more highly resolved scales, contact structure plays a more important role and contact clustering is potentially high. This clustering can be the result of spatial proximity [42] or common social contacts [43,44]. It creates a broad range of contact processes consistent with observed genetic and temporal information, and can make it difficult to estimate fundamental epidemiological parameters. Qualitative comparisons of patterns of contact identified through questionnaires or detailed investigation with the identified genetic sequences can reduce this uncertainty [45], although the transmission network for two different specific diseases can be substantially different from the social network that is common to both of them [9]. For example, in the case of a highly infectious disease such as the common cold or influenza, regular contact through shared use of public transport may be important but, for less-transmissible, longer-term illnesses such as tuberculosis, social networks (e.g., friendships) or migration patterns may predominate.

FMD is an important exemplar of forensic epidemiology. Outbreaks of disease in previously FMD-free countries are financially very costly and identifying the origin of infection can have important epidemiological, legal, and financial consequences. Such outbreaks have the advantage that they are usually monophyletic – in other words, all cases develop from a single introduction with diagnosis resulting in controls that prevent further introductions. Increasingly sophisticated methodologies have been applied to reconstructing transmission trees from the UK 2001 outbreak. Traditional epidemiological tracing [46] was followed by reconstructed trees using only spatial and temporal information [47]. Although next-generation sequencing was not available at the time of the epidemic, collection of viral samples across the epidemic has allowed retrospective analysis. Konig *et al.* [48] used consensus WGS to test specific hypotheses regarding the airborne spread of FMDV. Cottam *et al.* [30] used genetic data as a filter to identify a subset of the trees that were then ranked based on their likelihood given the associated space-time data. Further developments made possible the reconstruction of transmission trees and infection dates of susceptible premises, providing an example of a formal integration of genetic and spatiotemporal data within a single Bayesian inference scheme (Figure 2) [11]. Such joint inference schemes are a powerful (although often computationally intensive) approach to combining often disparate data (Box 2) and have been recently successfully applied to other RNA viruses [49,50].

Compared to RNA viruses such as FMDV, for more slowly replicating pathogens such as *Mycobacteria* spp. the epidemiological link between events may also be poorly resolved because the potentially infectious contacts are

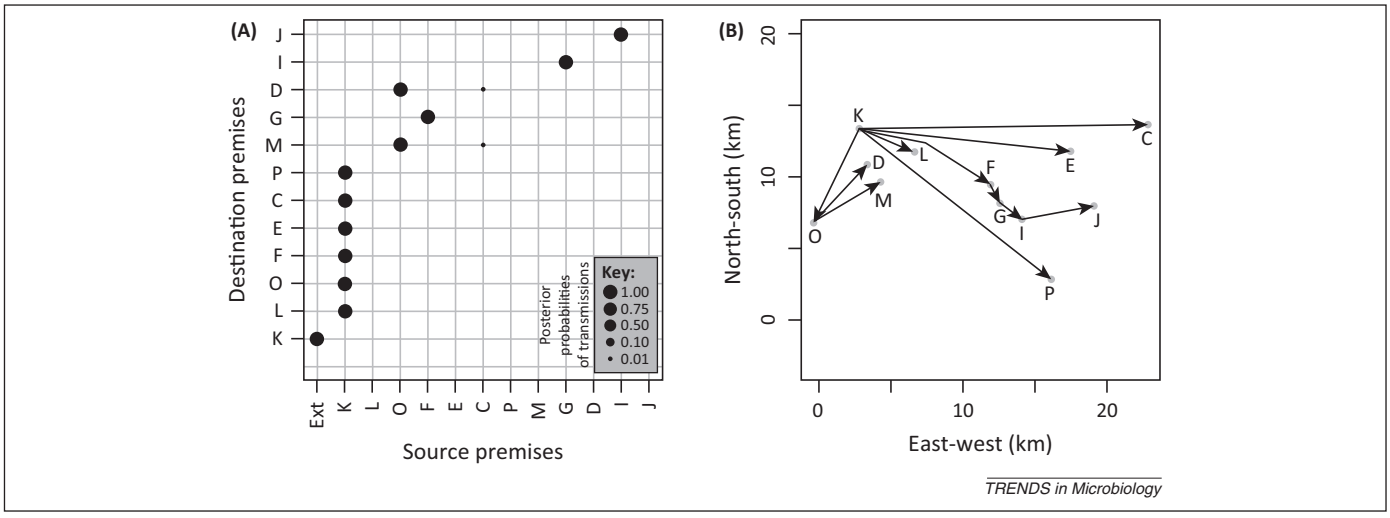

Figure 2. Phylodynamic reconstruction of a foot-and-mouth disease (FMD) epidemic. (A) Identified likelihood that a particular infected premises was the source of another infected premises based on a space–time–genetic model. Circle size is proportional to the relative likelihood of that event. (B) Spatial relationships among premises in the dataset. Reproduced from [11], with permission of the corresponding author.

less clearly defined than for sexually transmitted infections. For example, contacts may be relatively transient over the duration of the infectious period of the disease, although even in the case of HIV the impact of concurrency of relationships can be a challenging problem [51]. The utility of sequence information is further compromised when the relative mutation rate is also low, highlighting the importance of capturing as many mutations as possible while maintaining a low error rate. However, WGS has proven valuable in identifying, for example, evidence of the role of an immunocompromised drug-user in a single known transmission chain for *M. tuberculosis* [52] and in showing the existence of diverse lineages causing *Clostridium difficile* infections in hospitals [53]. One example where tracing is made difficult by prolonged incubation periods is for hospital infections of *Staphylococcus aureus* (SA). Although the underlying contact structure is well defined, SA is virtually omnipresent in the environment, and thus simple isolation of bacteria does not imply linkage to an outbreak [4,24]. Despite these issues, WGS for slowly evolving bacteria can be extremely valuable, especially when combined with detailed contact data [45].

Reservoirs and emerging infections

Multi-host pathogens present additional difficulties compared to pathogens that largely infect a single host species. Control options that may work in a single host species may not in another and, if the pathogen can persist in each host on its own, could render control ineffective. The evolutionary history of a multi-host pathogen system can be inferred by using discrete traits models in the same way as has been developed to identify transitions between geographical locations [54,55]). For example, these discrete traits methods can give estimates of when cross-species transmissions occurred [56]. In addition, the factors that influence the cross-species transmission rates can be inferred, such as genetic relatedness between species and species-range overlap [40].

These approaches are best applied where sampling is relatively even across species and where mutation rates in the different species are the same, or the differences in those rates are known. When sampled data are biased or unknown, or there are potentially significant differences in mutation rates, inference regarding which of the two species predominantly infects the other can be difficult. For example, the recovery of rabies isolates from domestic dogs is much more frequent than from wildlife, compromising our ability to use WGS alone to infer which is the source population [57]. Where mixing is limited, or equivalently where mixing between species is substantial, this will make inferring the role of the under-sampled host even more difficult (Figure 3). Where mixing is moderate, a much more distinctive signature can be obtained [56]. By contrast, the epidemiological signature of the hidden host is much clearer where mixing is high, especially in the case of intervention studies, where for example a reduction in the density of the under-sampled host may have a dramatic effect on disease incidence [58]. Thus when phylogenetic information can be combined with epidemiological data there is a potential to provide much deeper insight.

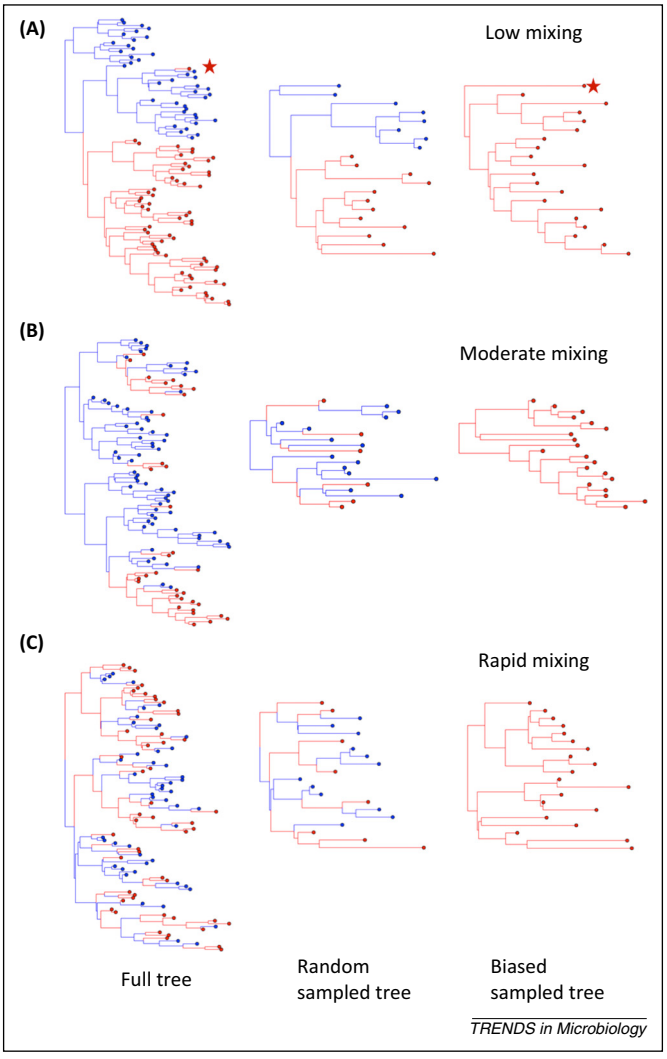

Figure 3. Biased sampling for multi-host systems causes problems for interpretation of genetic data even where the density of samples in one host is very high. The trees in the figure depict phylogenies of a pathogen in a two-host system; circles represent sampled sequences from the red or blue species. (A) For low mixing, random sampling reveals the relationship between the two host species but has a high probability of missing rare crossover events (red star). By contrast, dense sampling of one host (in red) will miss the existence of the second host species unless the crossover event is sampled, in which case the long branch length associated with it is instructive. By contrast, in (B) the distribution of branch lengths under biased sampling reveals the presence of unsampled events, although the nature of those events would not be determined by phylogenies alone. In (C), where mixing is substantial, the absence of data from the hidden host is likely unobserved or interpreted as greater variability in the mutation rate. It would be quickly revealed by even moderate sampling, although the phylogeny would remain difficult to distinguish from the case of a spillover host. The trees were created and displayed using a custom R script; random trees were created with the ape package, and a two-host discrete traits model was used with the package phytools to generate the ancestral and tip states.

The epidemic of bovine tuberculosis (bTB) in cattle in Britain and Ireland provides one example of that illustrates this potential. Epidemiological and demographic data from cattle are exceptionally well recorded, but with a relatively under-observed wildlife reservoir host, the Eurasian badger (*Meles meles*), whose role in maintaining the epidemic is as yet only partially quantified [59,60]. Use of WGS in a small recent study has shown the existence of a meaningful correlation between spatial distance and genetic distance at a kilometer scale, and this is poorly explained by livestock movements [61]. This spatial signature, although insufficient to identify direction, suggests

Box 3. WGS relevance to the pathogen emergence problems

Which is more important, multiple introductions or the presence of long chains, and how can we tell which is occurring? Understanding the latter question is both an important clue to answering the former and a diagnostic for when a host population of interest is likely to be vulnerable to an emergent infection. Possible reasons for case-clustering include multiple introductions from the same reservoir population, chains of emergent cases, and interactions between the two. Several steps are required for the emergence process to occur, such as exposure of the new host species to the emerging pathogen, spillover, and finally, sustained onward transmission. Pathogens introduced into new host populations can display variable values of R_0 (the average number of secondary infections that arise from one infected individual in a completely susceptible population). However, it is difficult to differentiate between clusters of cases caused by multiple introductions from those due to long transmission chains, particularly when host-switching pathogens circulate in species that share the same habitat (Figure I). In such cases WGS constitutes a valuable tool to complement epidemiological data because specific nucleotide polymorphisms could be used as genetic markers associated to the host.

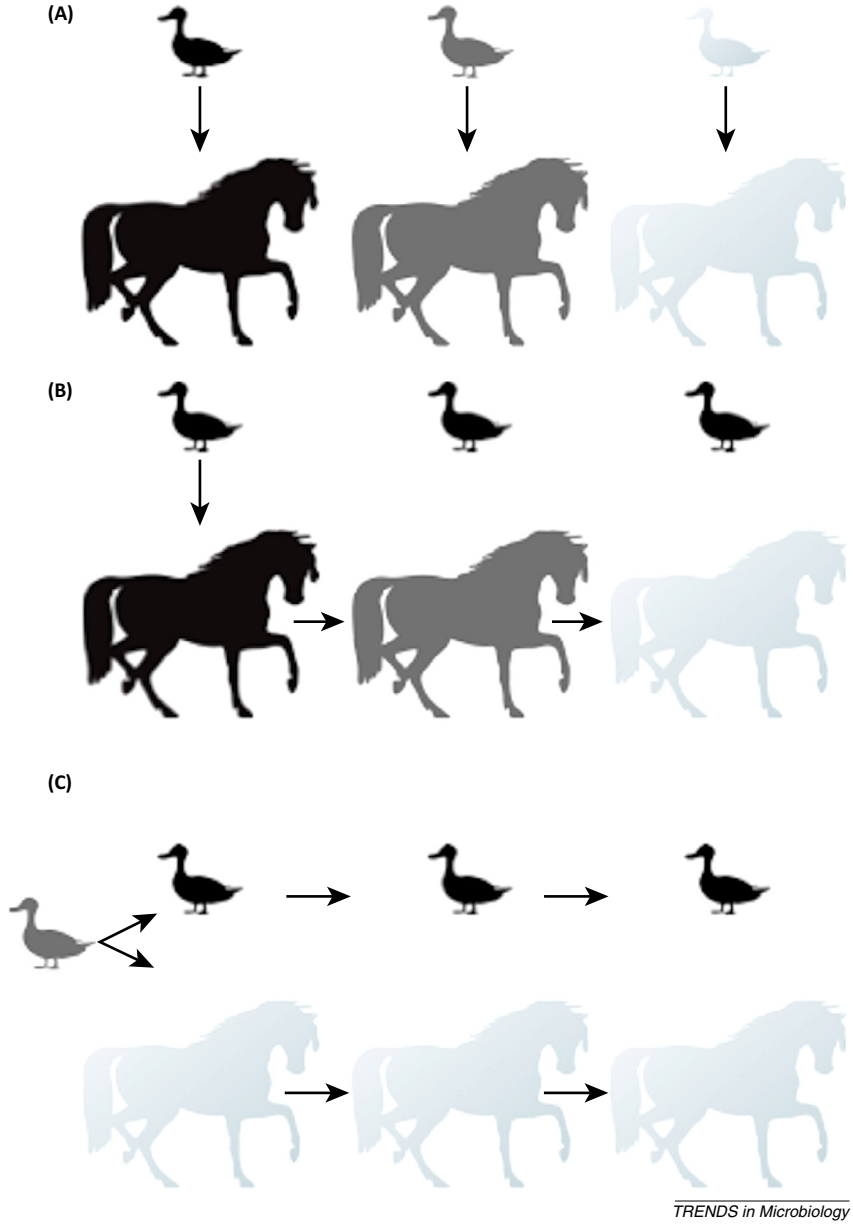

Figure I. Pathogen emergence or spillover? The figure represents the infection of horses by an avian pathogen. In **(A)**, black arrows represent an avian virus that is introduced multiple times (spillover) but cannot be transmitted among horses, whereas **(B)** represents a single introduction event followed by onward transmission. **(C)** Represents the circulation of two distinct lineages in both species that share a closely related ancestor. If differences between the lineages are minimal (such as a single nucleotide polymorphism), whole-genome sequencing (WGS) would be invaluable because otherwise there is likely to be no other detectable difference between the viruses in the two hosts.

that larger scale studies would be able to characterize the spatial–genetic relationship. This relationship is likely to be driven by a combination of cattle and badger activity in close spatial proximity, and its complexity suggest that mechanistic models will be needed to maximize the insight gained from the genetic data [61].

Emerging infections of humans and animals provide additional challenges (Box 3), and constitute a public health burden and a threat to food security and wildlife. Although emerging pathogens include bacteria, viruses, parasites, fungi, and even tumor cells [62,63], viruses are probably the most common source of emerging diseases [64]. Recent high-profile examples include the 2009 H1N1 influenza pandemic, the Middle East respiratory syndrome coronavirus (MERS-CoV) [65], and the H7N9 low-pathogenicity zoonotic avian influenza virus [66]. The process of pathogen emergence requires the introduction of the emerging pathogen into a susceptible population followed by onward transmission and, although the underpinning mechanisms of the entire process are as yet poorly understood, it is clear that ecological and evolutionary factors play crucial roles in it. The omics revolution, of which WGS is a one element, provides new tools to tackle the burden of emerging infectious diseases. For example, during outbreaks of disease the application of WGS has streamlined both pathogen identification as well as the likely source of infection, and even when unknown pathogens were involved [67].

Influenza A viruses (IAVs) constitute a textbook example of emerging viruses because they are well recognized for their ability to cross species barriers and establish in new host species. The genome of IAVs is segmented and thus coinfection of a single individual with multiple IAVs can lead to the generation of viruses that are different to the parental lineage through reassortment. Such viruses (reassortants) can have an expanded host range and/or different antigenic properties to which the susceptible population is naïve. In the past 100 years we have experienced four influenza pandemics: 'Spanish flu' in 1918 (an H1N1 virus), 'Asian flu' in 1957 (H2N2), 'Hong Kong flu' in 1968 (H3N2), and most recently 'swine flu' in 2009 (H1N1 again). During the latest pandemic two independent studies [68,69] generated the complete genome sequence of the pandemic virus and identified pigs as the source of infection nearly simultaneously with the World Health Organization (WHO) declaration of the pandemic in June 2009. These studies also showed the importance of epidemiological surveillance data to improve the accuracy of the reconstruction of the emergence process.

Differentiating between spillover (with limited transmission in the novel host) and emergence (where there are prolonged transmission chains) can be a challenging task, particularly when the donor and recipient species share the same ecological setting thus allowing multiple independent infections, or when the emerging pathogen has not yet acquired full transmissibility and its circulation is limited to short transmission chains (Box 3). For example in February 2013, a novel IAV (an H7N9 virus, referred to as LPAI H7N9) started infecting humans in Asia. The incidence was lower than for H1N1 in 2009, but with a much higher fatality rate, with 44 fatalities from 136 laboratory-confirmed cases diagnosed between February and October 2013. Thus far, current data clearly show that LPAI H7N9 has not acquired full transmissibility in humans and that chickens at live poultry markets are the most likely source of human infections [66]. However, extensive surveillance using WGS will be crucial to determining how H7N9 viruses circulate in nature, if they have (or will) become endemic in birds, and whether they will acquire mutations that could increase their transmissibility among humans.

Box 4. Outstanding questions

- Under what conditions does the inclusion of multiple sources of introduction (polyphyletic outbreaks) change the inference made in the reconstruction of transmission trees?
- Can robust, computationally tractable approaches be developed to handle partially observed outbreaks in which there are 'missing epidemiological links'?
- Can we develop general methods to use sequence data to estimate the proportion of cases that have been sampled from partially observed outbreaks?
- Can heterogeneities in susceptibility, transmissibility, or mode of transmission be robustly incorporated into phylodynamic models?
- How can these methods be adapted to take advantage of knowledge of minority variants within an individual or group as revealed by deep sequencing?
- Can we identify robust approaches that will allow us to choose between alternative models of transmission?

Another important source of genetic data that could reveal the origins of emerging viruses is archived material [70]. Sequencing viral genomes from frozen specimens has provided invaluable information about the pathogens that circulated in particular populations at a given time and in a particular location. However, obtaining WGS from pathogens derived from other sample types is not so straightforward.

With sequencing costs constantly decreasing, the application of WGS in routine epidemiological surveillance in the near future is likely to become common practice, with enormous knock-on benefits. Crucial to maximizing these benefits is recording ancillary data about both the transmission events and the underlying population.

Concluding remarks

Despite the already impressive list of achievements of WGS in epidemiology, there are many outstanding challenges remaining, particularly when considering endemic and less-intensively sampled situations (Box 4). Generating approaches to overcome these challenges will require the development of protocols for data collection and analytic tools that can only result from close interactions among clinicians, diagnosticians, epidemiologists, and mathematical biologists, highlighting the importance of transdisciplinary approaches to tackle the integration of all data sources. Importantly, these data sources will need to become broadly available across disciplines and for all legitimate research needs, while acknowledging the need for careful consideration of data protection and civil liberties issues.

Existing studies have demonstrated the resolution of WGS compared to previous typing schemes, highlighted the value of integrating diverse datasets, and demonstrated the insights gained from mathematical and statistical models. Technological advances now allow sequencing of

pathogen samples isolated from large numbers of infected individuals, often in epidemic real-time. Individual level denominator data from at-risk populations are now being collected, often on a daily basis, and also increasingly in real time. Advances in computing power now provide the engine for mathematical and statistical techniques by which disparate datasets can be analyzed. Although we are only now arriving at this point, the combination of big data and tractable analytical techniques provides the opportunity to transform our approach to controlling infectious diseases in both epidemic and endemic contexts, with WGS playing a leading role.

References

1 Ferguson, N.M. *et al.* (2001) The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* 292, 1155–1160
2 Parnell, S. *et al.* (2009) Optimal strategies for the eradication of asiatic citrus canker in heterogeneous host landscapes. *Phytopathology* 99, 1370–1376
3 Hampson, K. *et al.* (2009) Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biol.* 7, 462–471
4 Harris, S.R. *et al.* (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327, 469–474
5 Maiden, M.C. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145
6 Koser, C.U. *et al.* (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 8, e1002824
7 Dumitrescu, O. *et al.* (2011) Present and future automation in bacteriology. *Clin. Microbiol. Infect.* 17, 649–650
8 Roetzer, A. *et al.* (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med.* 10, e1001387
9 Kao, R.R. *et al.* (2007) Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *J. R. Soc. Interface* 4, 907–916
10 Bajardi, P. *et al.* (2012) Optimizing surveillance for livestock disease spreading through animal movements. *J. R. Soc. Interface* 9, 2814–2825
11 Morelli, M.J. *et al.* (2012) A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* 8, e1002768
12 Ypma, R.J. *et al.* (2012) Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. Biol. Sci.* 279, 444–450
13 Bhatt, S. *et al.* (2011) The genomic rate of molecular adaptation of the human influenza A virus. *Mol. Biol. Evol.* 28, 2443–2451
14 Morelli, M.J. *et al.* (2013) Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Vet. Res.* 44, 12
15 Nelson, M.I. and Holmes, E.C. (2007) The evolution of epidemic influenza. *Nat. Rev. Genet.* 8, 196–205
16 Lee, R.T. *et al.* (2010) All that glitters is not gold – founder effects complicate associations of flu mutations to disease severity. *Virol. J.* 7, 297
17 Holmes, E.C. (2009) The evolutionary genetics of emerging viruses. *Annu. Rev. Ecol. Evol. Syst.* 40, 353–372
18 Croucher, N.J. *et al.* (2013) Bacterial genomes in epidemiology – present and future. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 368, 20120202
19 Marttinen, P. *et al.* (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* 40, e6
20 Sanjuan, R. *et al.* (2010) Viral mutation rates. *J. Virol.* 84, 9733–9748
21 Lou, D.I. *et al.* (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19872–19877
22 Beerenwinkel, N. *et al.* (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3, 329
23 Bryant, J.M. *et al.* (2013) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect. Dis.* 13, 110
24 Walker, T.M. *et al.* (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* 13, 137–146
25 Streicker, D.G. *et al.* (2012) Rates of viral evolution are linked to host geography in bat rabies. *PLoS Pathog.* 8, e1002720
26 Stack, J.C. *et al.* (2013) Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. Biol. Sci.* 280, 20122173
27 Ypma, R.J. *et al.* (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195, 1055–1062
28 Onnela, J.P. *et al.* (2011) Geographic constraints on social network groups. *PLoS ONE* 6, e16939
29 Volz, E.M. *et al.* (2012) Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Comput. Biol.* 8, e1002552
30 Cottam, E.M. *et al.* (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Biol. Sci.* 275, 887–895
31 Zaas, A.K. *et al.* (2013) A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. *Sci. Transl. Med.* 5, 203ra126
32 Savill, N.J. *et al.* (2007) Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *J. R. Soc. Interface* 4, 235–241
33 Onnela, J.P. *et al.* (2007) Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7332–7336
34 Hufnagel, L. *et al.* (2004) Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15124–15129
35 Bajardi, P. *et al.* (2011) Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE* 6, e16591
36 Chis Ster, I. *et al.* (2012) Within-farm transmission dynamics of foot and mouth disease as revealed by the 2001 epidemic in Great Britain. *Epidemics* 4, 158–169
37 Templeton, A.R. (2009) Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Mol. Ecol.* 18, 319–331
38 Hedge, J. *et al.* (2013) Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biol. Lett.* 9, 20130331
39 Hughes, J. *et al.* (2012) Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog.* 8, e1003081
40 Faria, N.R. *et al.* (2011) Toward a quantitative understanding of viral phylogeography. *Curr. Opin. Virol.* 1, 423–429
41 Volz, E.M. *et al.* (2009) Phylodynamics of infectious disease epidemics. *Genetics* 183, 1421–1430
42 Keeling, M.J. (1999) The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B: Biol. Sci.* 266, 859–867
43 Girvan, M. and Newman, M.E. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826
44 Handcock, M.S. *et al.* (2007) Model-based clustering for social networks. *J. R. Stat. Soc. Ser. A* 170, 301–354
45 Gardy, J.L. *et al.* (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* 364, 730–739
46 Gibbens, J.C. *et al.* (2001) Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Vet. Rec.* 149, 729–743
47 Haydon, D.T. *et al.* (2003) The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. R. Soc. Lond. B: Biol. Sci.* 270, 121–127
48 Konig, G.A. *et al.* (2009) Sequence data and evidence of possible airborne spread in the 2001 foot-and-mouth disease epidemic in the UK. *Vet. Rec.* 165, 410–411
49 Jombart, T. *et al.* (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* 10, e1003457

50 Mollentze, N. *et al.* (2013) A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B* 3251 http://dx.doi.org/10.1098/rspb.2013.3251
51 Morris, M. and Kretzschmar, M. (1997) Concurrent partnerships and the spread of HIV. *AIDS* 11, 641–648
52 Schürch, A.C. *et al.* (2010) The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. *Infect. Genet. Evol.* 10, 108–114
53 Eyre, D.W. *et al.* (2013) Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N. Engl. J. Med.* 369, 1195–1205
54 Lemey, P. *et al.* (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5, e1000520
55 Rabaa, M.A. *et al.* (2010) Phylogeography of recently emerged DENV-2 in southern Viet Nam. *PLoS Negl. Trop. Dis.* 4, e766
56 Mather, A.E. *et al.* (2013) Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. *Science* 341, 1514–1517
57 Lembo, T. *et al.* (2008) Exploring reservoir dynamics: a case study of rabies in the Serengeti ecosystem. *J. Appl. Ecol.* 45, 1246–1257
58 Woodroffe, R. *et al.* (2009) Bovine tuberculosis in cattle and badgers in localized culling areas. *J. Wildl. Dis.* 45, 128–143
59 Godfray, H.C. *et al.* (2013) A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain. *Proc. Biol. Sci.* 280, 20131634
60 Donnelly, C.A. and Nouvellet, P. (2013) The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England. *PLoS Curr. Outbreaks* http://dx.doi.org/10.1371/currents.outbreaks.097a904d3f3619db2fe78d24bc776098
61 Biek, R. *et al.* (2012) Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. *PLoS Pathog.* 8, e1003008
62 McCallum, H. *et al.* (2009) Transmission dynamics of Tasmanian devil facial tumor disease may lead to disease-induced extinction. *Ecology* 90, 3379–3392
63 Pearse, A.M. and Swift, K. (2006) Allograft theory: transmission of devil facial-tumour disease. *Nature* 439, 549
64 Jones, K.E. *et al.* (2008) Global trends in emerging infectious diseases. *Nature* 451, 990–993
65 Zaki, A.M. *et al.* (2012) Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* 367, 1814–1820
66 Chen, Y. *et al.* (2013) Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome. *Lancet* 381, 1916–1925
67 Chandriani, S. *et al.* (2013) Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1407–E1415
68 Smith, G.J. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459, 1122–1125
69 Garten, R.J. *et al.* (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* 325, 197–201
70 Wright, C.F. *et al.* (2013) Reconstructing the origin and transmission dynamics of the 1967-68 foot-and-mouth disease epidemic in the United Kingdom. *Infect. Genet. Evol.* 20C, 230–238
71 Pybus, O.G. and Rambaut, A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10, 540–550
72 Lemey, P. *et al.* (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* 27, 1877–1885
73 Drummond, A.J. *et al.* (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161, 1307–1320
74 Pybus, O.G. *et al.* (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15066–15071
75 Dearlove, B. and Wilson, D.J. (2013) Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philosophical transactions of the Royal Society of London. Series B. *Biol. Sci.* 368, 20120314
76 Volz, E.M. and Frost, S.D. (2013) Inferring the source of transmission with phylogenetic data. *PLoS Comput. Biol.* 9, e1003397
77 Stadler, T. *et al.* (2013) Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. U.S.A.* 110, 228–233