Genome Biology

REVIEW

Genomic analysis of emerging pathogens: methods, application and future trends

Lucy M Li*, Nicholas C Grassly and Christophe Fraser

Abstract

The number of emerging infectious diseases is increasing. Characterizing novel or re-emerging infections is aided by the availability of pathogen genomes. In this review, we evaluate methods that exploit pathogen sequences and the contribution of genomic analysis to understand the epidemiology of recently emerged infectious diseases.

Introduction

When a pathogen crosses over from animals to humans, or an existing human disease suddenly increases in incidence, the infectious disease is said to be 'emerging'. The number of emerging infectious diseases (EIDs) has increased over the last few decades, driven by both anthropogenic and environmental factors [1]. These include the expansion of agricultural land, which increases the exposure of livestock and humans to infections in wildlife [2]; a greater volume of air traffic, enabling EIDs to rapidly spread across the world [3,4]; and climate change, which alters the ecology and density of animal vectors, thereby introducing diseases to new geographic locations [5]. Novel strains of existing pathogens also have the potential to cause large epidemics. The over- and misuse of antimicrobial drugs have contributed to the growing number of drug-resistant pathogen strains [6,7].

Detecting, characterizing and responding to an EID requires co-ordination and collaboration between multiple sectors and disciplines. Laboratory-based research helps to characterize the pathogen and its interactions with host cells, but is less useful for quantitative understanding of population-level disease dynamics. Modeling approaches enable a large number of hypotheses to be tested, which might not be logistically or ethically feasible in laboratory and field experiments. In addition to characterizing past disease dynamics, modeling future trends informs decisions regarding outbreak response and resource allocation [8]. Modeling plays an especially important role in epidemiological studies of infectious disease spread, because the transmission of infectious disease between individuals is not directly observable. At the individual level, transmission times and who infected whom are typically unknown. And at the population level, disease burden needs to be inferred from observable data. Important public health questions such as how quickly an epidemic spreads and how many people will be infected are hard to quantify without a mechanistic understanding of underlying factors driving disease transmission. By expressing disease spread in mathematical terms, statistical properties of epidemics can be estimated to help address specific questions regarding disease spread and control efforts [9].

Another discipline contributing to the study of EIDs is pathogen genomics. As sequencing technology has become more accessible and affordable, genetic analysis has played an increasingly important role in infectious disease research. Sequencing pathogens can confirm suspected cases of an infectious disease, discriminate between different strains, and classify novel pathogens. In addition to examining individual pathogen sequences, multiple sequences can be analyzed together using phylogenetic methods to elucidate evolutionary [10] and transmission [11] history. Just as mathematical models of disease transmission help to capture the epidemiological properties of an infectious disease, modeling the molecular evolution of pathogen genomes is important for phylogenetic methods.

Besides characterizing the genetics and evolution of a pathogen, mathematical models used in population genetics link demographic and evolutionary processes to temporal changes in population-level genetic diversity. The coalescent population genetics framework was developed so that demographic history could be inferred from the shape of the genealogy linking sampled individuals [12,13]. More recently, the birth-death model has been applied to infectious diseases to infer epidemiological history from a genealogy [14,15]. Given the link between pathogen

* Correspondence: mengqi.li09@imperial.ac.uk
Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, London W2 1PG, UK

evolution and disease transmission, there is a trend towards integrating both epidemiologic and genetic data in the same analytical framework [16-18].

In this review, we provide an overview of recent developments in genomic methods in the context of infectious diseases, evaluate integrative methods that incorporate genetic data in epidemiological analysis, and discuss the application of these methods to EIDs.

Role of genetics in studying infectious diseases

Over the last two decades, sequence data have increased in quality, length and volume due to improvements in the underlying technology and decreasing costs. As a result, pathogen sequences are regularly collected during routine surveillance and clinical studies. Just as mathematical modeling can be used to analyze surveillance data to reveal details of disease transmission (Box 1), analysis of pathogen genomes employs mathematical frameworks to elucidate pathogen biology, evolution and ecology (Figure 1).

At the most basic level, mathematical models are used to find the optimal alignment of pathogen sequences. Multiple sequence alignment is useful for finding highly conserved or variable regions, shedding light on the molecular biology of the pathogen. Furthermore, coupling sequences with clinical information can help identify the contribution of polymorphic sites to disease. Revealing the evolutionary history of a pathogen requires a quantitative description of relatedness. Based on polymorphic sites in the sequence alignment, a model of sequence evolution is then used to reconstruct the phylogeny [19]. Often, there is insufficient genetic diversity in the sample to fully infer the phylogeny without ambiguity. In such a case, it is useful to consider a tree as an unknown set of

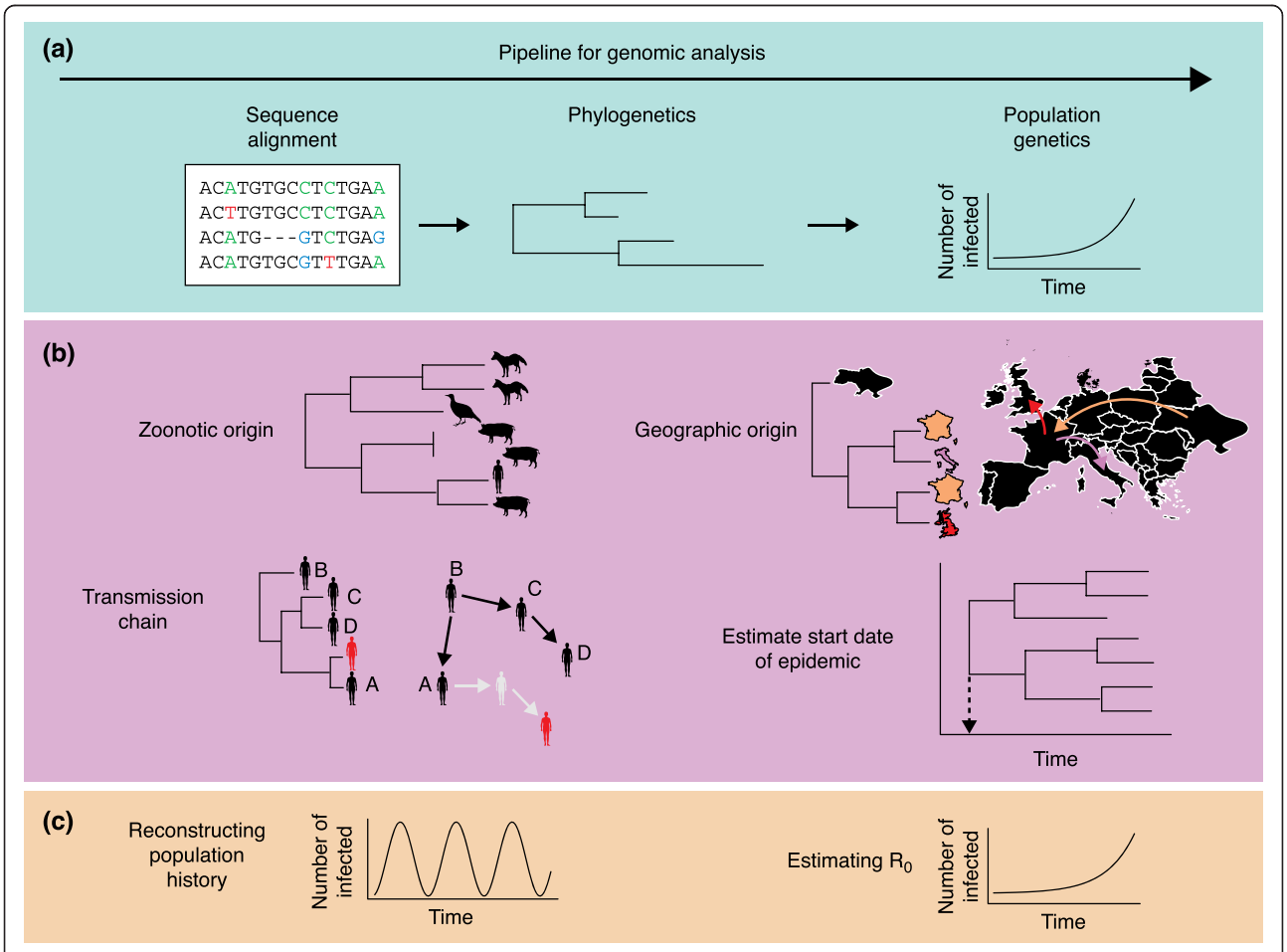

Figure 1 Contribution of genomic analysis to epidemiological studies of emerging infectious diseases. (a) Genomic analysis begins with obtaining a multiple sequence alignment of pathogen sequences from which a phylogeny can be built to represent the evolutionary relationship between samples. Further population genetic analysis using the coalescent framework can reveal the population history of the pathogen based on the sample phylogeny. **(b)** Coupling phylogeny with additional information is useful for uncovering zoonotic origins, the spatiotemporal patterns of disease spread, and transmission chains. The results of such phylogenetic analysis should be interpreted with care as the direction of transmission is not always clear and there might exist missing intermediate links. **(c)** Coalescent analysis of pathogen genealogy is used to characterize past epidemiological dynamics and estimate epidemiological parameters, such as the reproductive number.

parameters and obtain its posterior probability distribution using a Bayesian framework, such as the Markov Chain Monte Carlo (MCMC) approaches [20,21].

Biological samples from which pathogen genetic material is sequenced are usually associated with geographic or temporal information (Figure 1b). When this additional information is available, phylogenetic methods can reveal the spatiotemporal spread of the pathogen in the population. If an outbreak is densely sampled, then the pathogen phylogeny provides information about the underlying transmission network and helps to uncover who infected whom [22,23], though phylogenetic clustering alone is usually not sufficient to prove direct transmission or direction of infection (Figure 1b).

Incorporating sampling times helps to convert a phylogeny specified in units of nucleotide substitutions to a phylogeny specified in units of time [24]. The conversion is straightforward if sequence evolution follows a strict molecular clock, whereby the rate of substitution remains constant over time. However, selection pressure and population bottlenecks can lead to changes in the rate of substitution [25]. More flexible models have been developed to incorporate time-varying rates of evolution [26,27]. With branch lengths in units of real time, the start date of an epidemic can be estimated. Whereas phylogenetics aims to delineate the relationship between individuals, population genetics aims to link population processes to observed patterns of genetic diversity. Inferences regarding pathogen population history are based on the genealogy, or ancestry, of sequences from sampled individuals, and often carried out in a retrospective population genetics framework known as the coalescent [12] (Box 2). A genealogy describes the ancestry of sampled individuals. Going backwards in time, pairs of lineages coalesce when they share a common ancestor, until the last two lineages coalesce at the time of the most recent common ancestor (TMRCA) for the entire sample.

Since the turn of the century, the coalescent has been increasingly applied to infectious disease research to infer epidemic history from pathogen sequences, thereby linking pathogen evolutionary history to disease epidemiology (Figure 1c). The method is especially useful for analyzing infectious diseases with mild or asymptomatic infections, for which case-based surveillance data severely underestimate prevalence, because the coalescent assumes a small sample compared to the population size [28-30].

Other approaches have been developed to make epidemiological inferences from genetic data. Of particular note is the birth-death model [31], which describes the rates of transmissions, recoveries and deaths, and sampling events in terms of the sample genealogy [14]. Just as there are coalescent methods incorporating population structure [32-34] and compartmental models [35-37], similar methods exist in the birth-death framework [38,39]. Unlike the coalescent framework, the birth-death model is still valid for densely sampled populations, which makes it more useful for studying small outbreaks. However, accurately inferring epidemiological parameters depends on correctly specified sampling proportions [40]. Although the two approaches are methodologically different, both aim to reconstruct pathogen population history and produce estimates of epidemiological parameters, such as the reproductive number (R_0). The focus on the coalescent framework in this review is due to its more pervasive use in the literature and its greater versatility when integrated with epidemiological models compared to birth-death models.

Because of the simplistic assumptions of population genetics models, the population size inferred using coalescent-based methods cannot be directly interpreted as pathogen population size (prevalence of infection). It is rather the effective population size, N_e (Box 2), which refers to the size of a Wright-Fisher population that would produce the same level of genetic diversity as observed in the sample. In real populations, the variance of the offspring distribution (Box 1) is higher than expected in a Wright-Fisher population due to heterogeneity in host infectiousness, non-random mixing of the population, and migration events. The consequence of a large variance is that there is a greater discrepancy between the effective and census population sizes [41]. Accounting for the dispersion of the offspring distribution is especially important when analyzing infectious disease data because of the widespread occurrence of transmission heterogeneity [42].

Another statistical property of epidemics affecting the results of modeling studies is the generation time distribution, which describes the time between infection of the primary case and of secondary cases. Obtaining an estimate of the generation time is important for two reasons. First, estimates of R_0 from the initial growth rate of an epidemic depend on the generation time distribution [43]. As R_0 is the mean of the offspring distribution, its value affects the relationship between the effective population size, N_e, and the census population size, N. Second, the coalescent model was originally specified in units of generations, and so estimates in this framework need to be converted to natural units using the generation time, T_g.

Because transmission events are rarely observed, the generation time distribution is often approximated by the distribution of the serial interval, which is the time between onset of symptoms in the primary and secondary cases. The two distributions generally share the same mean but might have different variances [44]. Furthermore, the observed generation time decreases as the epidemic grows but increases again after the epidemic peak due to right censoring [45].

Integrating genetics with other data

As both sequence and surveillance data contain information regarding the transmission process, simultaneously analyzing both datasets should yield more accurate estimates of epidemiological parameters than separate analyses [17]. The recently established discipline of phylodynamics takes an interdisciplinary approach to understand the pathogen phylogenetics and epidemiology in terms of disease transmission.

Most efforts thus far have focused on enhancing phylogenetic and population genetic analyses by incorporating spatial and temporal information about the sequences. The molecular clock model assumes a constant rate of evolution and thus helps to estimate the time of the most recent common ancestor of the sample, which approximates the start date of an epidemic. Molecular clock analysis has been used to date the emergence of a range of emerging pathogens from HIV [46] to multidrug-resistant *Streptococcus pneumoniae* [47].

Linking geographic information with sequences can reveal the spatial spread of infectious disease. Phylogenetic reconstruction of seasonal influenza (H3N2) sequences has revealed the contribution of viral circulation in temperate regions to the global genetic diversity of influenza, and determined that not all epidemics in temperate regions are seeded by strains from South East Asia [48,49]. Also using global sequences, hepatitis C virus (HCV) subtypes were shown to spread from developed to developing countries [50]. Finally, phylogeographic analysis of methicillin-resistant *Staphylococcus aureus* samples identified England as the source of the EMRSA-15 lineage [51].

By contrast, there have been relatively few studies incorporating genetic data into epidemiological frameworks. Although genetic analysis plays an important role in elucidating transmission links in disease outbreaks [20,21,52], its integration with epidemiological models to understand population-level disease dynamics has been more limited. In one of the first papers to link coalescent inference to mathematical models in epidemiology, the effective population sizes of HIV-1 subtypes A and B were estimated from the maximum likelihood trees of viral sequences [53]. In addition to revealing population sizes, Pybus *et al*. [54] estimated the R_0 values of HCV subtypes (1a, 1b, 4 and 6) by inferring the epidemic growth rate from viral genealogy. Taking integration a step further, the coalescent process has been described for compartmental epidemiological models such as the Susceptible-Infected-Recovered (SIR) model, thereby enabling epidemiological parameters to be inferred from the genealogy [35]. To infer demographic history from both pathogen genomes and epidemiological data, Rasmussen *et al*. [17] developed a Markovian framework in which the population size at each time step was estimated by taking into account both the surveillance data and the genealogy. The epidemic history reconstructed using both datasets was more accurate than when analyzing each type of data separately.

In all the above methods, the genealogy of the sampled sequences was fixed. However, there might be great uncertainty regarding the order and the timing of coalescence, especially if the sequences are sampled within a short time period. While genealogical reconstruction using Bayesian MCMC approaches allows phylogenetic uncertainty to be incorporated into estimates of population size [13,31], an integrative model is lacking in which uncertainties arising from both genetic and epidemiological data are incorporated during demographic reconstruction.

Application to emerging pathogens

Models of pathogen evolution and mechanistic models of disease spread have increased in complexity. There is also greater computational power to test these models with data. However, these sophisticated models have mostly been applied to infectious diseases for which abundant data are available. For example, new methods are most often tested on the HIV-1 pandemic [15,34,35,55], for which data have been extensively collected from various settings and sources since the virus was first characterized three decades ago. It is worthwhile to evaluate how genomic methods have been applied to other diseases that have emerged more recently. In this section, we will present three case studies of recently emerged infectious diseases to illustrate the power and shortcomings of genomic methods discussed in this review.

Ebola virus emergence in West Africa

Since emerging in Guinea in March 2014, Ebola virus (EBOV) has spread to other countries in Western Africa, resulting in the largest outbreak of Ebola since it was first identified in 1976. The first viral genomes were made available just a month after alarm was raised about a new Ebola outbreak in Guinea [56], with further sequences collected in Sierra Leone [57]. By aligning all the genomes, a number of polymorphic sites were identified, including eight in highly conserved regions of the genome. Further association studies are needed to clarify the role of these genetic variants in determining disease outcome. Using the sampling dates of the sequences and a molecular clock model, phylogenetic analysis of 81 EBOV sequences revealed a start date of February 2014 in Guinea, spreading to Sierra Leone by April 2014 [57].

Uncovering the relationship between the 2014 EBOV lineage and previous EBOV outbreaks has proved trickier than understanding the disease dynamics during the 2014 outbreak. Initial phylogenetic analysis suggested that lineages causing the present outbreak did not cluster with EBOV strains that caused earlier outbreaks in Central Africa [56]. However, Dudas and Rambaut [58]

noted that the divergence of Guinea sequences from those of previous outbreaks was because they were sequenced most recently and had accumulated the highest number of substitutions. Assuming that the EBOV genome followed a molecular clock model, the authors re-rooted the tree to a lineage that caused an outbreak in 1976 [58]. Instead of silently circulating in West Africa, the EBOV lineage causing the current outbreak likely descended from a lineage that previously caused outbreaks in the Democratic Republic of Congo.

These studies highlight two issues. First, correct rooting of a phylogeny is important for accurate inference of past epidemic history. Correct rooting can be achieved by using an out-group, but one was not available in the case of this EBOV strain. This leads onto the second issue. Without sequences from animal hosts, the mechanism by which EBOV was sustained between outbreaks remains unknown.

Middle East respiratory syndrome coronavirus

Middle East respiratory syndrome coronavirus (MERS-CoV) first appeared in Saudi Arabia in 2012, and has since been reported in several neighboring countries in the Arabian Peninsula and on other continents [59].

Despite the dearth of sequence data, coalescent-based analysis of 10 genomic sequences produced estimates of the TMRCA (March 2012; 95% confidence interval (CI): November 2011 to June 2012), R_0 (1.21; 95% CI: 1.08, 1.40), and doubling time (43 days; 95% CI: 23, 104 days) [60]. Without further sequencing of the animal reservoirs, the authors could not infer whether these estimates applied to the animal reservoir or the human epidemic, because the methods are agnostic as to where transmission and evolution occur. The credible intervals around the estimates were unsurprisingly large given the small sample size.

Unlike the 2014 EBOV outbreak, which is sustained by human-to-human transmission [57], there appears to have been multiple introductions of MERS-CoV into the human population. Identification of the animal reservoir is therefore crucial for establishing risk factors of infection and planning appropriate interventions to control the disease. Since bats are reservoirs for other coronaviruses, their being a reservoir host is possible. A 182-nucleotide-long region of the RNA-dependent RNA polymerase gene was found to be 100% identical between a viral sample from a patient in Saudi Arabia and from a bat nearby, though the region is known to be highly conserved [61]. However, antibodies against human MERS-CoV have been detected in dromedary camels [62], the camel MERS-CoV genome is similar to human MERS-CoV [62], and there are reports of close contact between patients and camels [63]. Phylogenetic analysis of coronavirus sequences from bats, dromedaries and humans indicate a bat origin, with dromedary camel as an intermediate host [64]. It is possible that there are other animal reservoirs not yet sampled, which highlights the need to carry out extensive animal surveillance to characterize the emergence of an infection in humans.

Unraveling the complex evolutionary history of pandemic H1N1 influenza

With sequences collected over three decades from humans, pigs and birds, the origin of the pandemic H1N1 influenza A strain (pdmH1N1 or 'swine flu') was elucidated soon after emergence. Within two months of the first reported case of swine flu in humans, genomic analysis of the novel influenza strain had been carried out. A phylogeny was constructed for each of the eight genomic segments with sequences from humans, swine and birds. Comparison of these eight phylogenies revealed a complex history of reassortment with a mixture of gene segments from all three groups. The start of the pandemic was estimated to be the end of 2008 or early 2009, and the dates of the reassortment events leading to pdmH1N1 were also obtained [10]. Without good surveillance of influenza in the animal reservoir, the origin of the novel strain would have been difficult to uncover.

By analyzing 11 hemagglutinin sequences collected over a one-month period, the start date of the epidemic was estimated to be in late January 2009 [65]. Repeating the phylogenetic and molecular clock analyses with a further 12 sequences shifted the estimated start date two weeks earlier. Fitting an exponential growth model to the sequence data, R_0 was estimated to be 1.22, slightly lower than inferred from epidemiological data but with overlapping confidence intervals.

To determine at which point during the pandemic coalescent analysis would have provided accurate and precise estimates of evolutionary rate, R_0 and TMRCA, real-time estimates of these parameters were obtained for genomic sequences collected in North America [66]. Accurate estimates could have been obtained as early as May, when 100 viral genomes had been sequenced. More precise estimates could have been obtained by the end of June, when 164 had been sequenced. However, inclusion of more sequences of longer length only slightly improved the accuracy of initial estimates [66].

Future directions

Most statistical models in population genetics have focused on the application of such methods to viruses, although this bias is perhaps unsurprising given the large proportion of EIDs caused by viruses [1]. Whole-genome sequencing of bacterial isolates is becoming more widespread, and can help to uncover genetic determinants of clinical severity, elucidate pathogen-host interactions, and

quantify evolutionary rates at within- and between-host levels [67]. Epidemiological investigations using bacterial genomes have also been possible. Even though bacteria acquire point mutations at a lower rate per base than viruses, longer bacterial genomes have provided sufficient genetic resolution for phylogenetic analysis. For example, whole-genome sequencing has been used to refine the tuberculosis transmission network built using contact information [21], and to investigate an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital and surrounding community in near real-time [68]. The need for longer sequences when conducting epidemiological studies of bacterial infections adds to the per-sample cost of sequencing, and more computational resources are required for coalescent-based inference of pathogen history. However, this latter limitation may be overcome by only analyzing polymorphic sites if samples are similar.

Demographic reconstruction of emerging bacterial pathogens using coalescent-based approaches has been limited compared to work on viral pathogens. In one such study, the temporal changes in genetic diversity of *Streptococcus pneumoniae* in Iceland were estimated based on the coalescent model [47]. This study was limited to a single multidrug-resistant lineage in a single location, with data collected over decades. Over longer evolutionary time-scales, the accumulation of diversity through recombination can obscure phylogenetic relationships. More complex evolutionary models would be required to taken into account these genomic changes, increasing the uncertainty surrounding demographic estimates from genomic data.

In addition to performing analyses with longer sequences, there is also a need to develop methods that exploit as many sequences in the sample as possible. For population studies, available sequences are often subsampled to remove individuals from the same household or in the same close contact network to have a representative sample of the population. Furthermore, sequences from the same individuals are often discarded, though these may be informative for within-host evolution. Although some effort has been made to link within-host to between-host evolution [52,69], the effect of within-host evolution on population genetic inference is still not well studied. Combining analyses across different scales could improve the accuracy of epidemiological predictions and provide better mechanistic explanations of observed trends.

Conclusion

Genomic studies have contributed to better understanding of EIDs and their spatiotemporal spread. Sophisticated statistical methods have been developed to uncover the epidemiological features of infectious diseases based on the genealogy of their sequences. There is also growing

Box 1. Key concepts in mathematical modeling of infectious disease transmission

Representing infectious disease transmission in a mathematical framework requires distilling complex observations into simple but informative expressions. Perhaps the most important statistical property of interest to an epidemiologist is the basic reproductive number, R_0, which represents the mean number of secondary infections caused by each infected individual in a wholly susceptible population. An epidemic can only occur if $R_0 > 1$. As an epidemic progresses, or if there is pre-existing immunity in a population, R_0 is no longer appropriate for describing the number of secondary infections per primary infection. Instead the effective reproductive number, R, is used. Another important statistical property of an epidemic is the generation time, T_g, which is the mean time between when an individual becomes infected and when they infect others. The combination of R_0 and T_g provides an indication of how quickly an epidemic will spread.

The most common type of model used in infectious disease research is the compartmental model. Given a set of parameters, a compartmental model tracks the temporal dynamics of subpopulations that are characterized by disease status. For example, a Susceptible-Infected-Recovered (SIR) model describes the changes in the number of susceptible, infected and recovered (and immune) individuals. R_0 can be calculated by inferring the set of model parameters that can generate the epidemiological dynamics most similar to those observed in the data. Increasingly, model parameters are inferred in a Bayesian framework. Bayesian inference finds the posterior probability distribution of parameters, given prior information and the data. Exploring all possible parameter combinations is intractable. The use of Markov Chain Monte Carlo (MCMC) for Bayesian statistical inference has enabled efficient estimation of the posterior probability distribution when the distribution cannot be computed analytically [70].

Obtaining estimates of R_0 and T_g is not always sufficient to predict epidemic trajectory if there is significant heterogeneity between individuals. The offspring distribution with mean R and variance σ^2 describes the probability distribution of the number of secondary infections caused by each infected individual. In compartmental models, the offspring distribution is not explicitly specified but follows from the specification of the model - in the case of the SIR model it follows a geometric distribution. For certain diseases, the offspring distribution is more dispersed than captured by the geometric distribution [42]. In other words, most individuals cause no further infections whereas a few individuals are super-spreaders who cause the majority of infections. Accurate estimate of σ^2 is important for predicting epidemic outcome and assessing control measures.

Box 2. Coalescent inference from genetic data

Just as compartmental models can be fitted to surveillance data to infer the epidemiological dynamics of an infectious disease (Box 1), the coalescent framework allows inference of population history from pathogen sequences. The coalescent model describes the statistical properties of the genealogy underlying a small sample of individuals from a large population. In the simplest case, the forward-time dynamics of the population is assumed to follow the Wright-Fisher model, in which the haploid population has discrete, non-overlapping generations, undergoes neutral evolution, and remains the same size [71,72]. Extensions to the coalescent have assumed more complex population dynamics described by deterministic population equations [73], compartmental disease models [35], or non-parametric approaches [13,55,74,75].

Within this framework, going backwards in time, individuals in the current generation are randomly assigned to parents in the previous generation. If two individuals have the same parent, then a coalescent event has occurred. Eventually, all lineages in the sample coalesce to a single individual known as the most recent common ancestor of the sample.

The rate of coalescence is inversely related to population size. If the population follows the Wright-Fisher model, evolutionary changes are selectively neutral, so the shape of the genealogy reflects only demographic changes.

effort to integrate genomic analysis with analysis of epidemiological data. In recent cases of EIDs, genomic data have helped to classify and characterize the pathogen, uncover the population history of the disease, and produce estimates of epidemiological parameters.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Nick Croucher for discussions on bacterial genomics. LL is funded by a Medical Research Council Doctoral Training Partnership Studentship.

Published online: 22 November 2014

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P: **Global trends in emerging infectious diseases.** *Nature* 2008, **451:**990–993.
2. Pulliam JR, Epstein JH, Dushoff J, Rahman SA, Bunning M, Jamaluddin AA, Hyatt AD, Field HE, Dobson AP, Daszak P: **Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis.** *J R Soc Interface* 2012, **9:**89–101.
3. Wilder-Smith A, Gubler DJ: **Geographic expansion of dengue: the impact of international travel.** *Med Clin North Am* 2008, **92:**1377–1390.
4. Khan K, Arino J, Hu W, Raposo P, Sears J, Calderon F, Heidebrecht C, Macdonald M, Liauw J, Chan A, Gardam M: **Spread of a novel influenza A (H1N1) virus via global airline transportation.** *New Engl J Med* 2009, **361:**212–214.
5. Le Guenno B, Camprasse MA, Guilbaut JC, Lanoux P, Hoen B: **Hantavirus epidemic in Europe, 1993.** *Lancet* 1994, **343:**114–115.
6. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, Ziazarifi AH, Hoffner SE: **Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran.** *Chest* 2009, **136:**420–425.
7. Ohnishi M, Golparian D, Shimuta K, Saika T, Hoshina S, Iwasaku K, Nakayama S, Kitawaki J, Unemo M: **Is *Neisseria gonorrhoeae* initiating a future era of untreatable gonorrhea?: Detailed characterization of the first strain with high-level resistance to ceftriaxone.** *Antimicrob Agents Chemother* 2011, **55:**3538–3545.
8. Anderson RM, May RM: *Infectious Diseases of Humans: Dynamics and Control*, Volume 28. Oxford: Oxford University Press; 1991.
9. Grassly NC, Fraser C: **Mathematical models of infectious disease transmission.** *Nat Rev Microbiol* 2008, **6:**477–487.
10. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JS, Guan Y, Rambaut A: **Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.** *Nature* 2009, **459:**1122–1125.
11. Cottam EM, Haydon DT, Paton DJ, Gloster J, Wilesmith JW, Ferris NP, Hutchings GH, King DP: **Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001.** *J Virology* 2006, **80:**11274–11282.
12. Kingman JF: **On the genealogy of large populations.** *J Appl Probability* 1982, **19:**27–43.
13. Drummond AJ, Rambaut A, Shapiro B, Pybus OG: **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005, **22:**1185–1192.
14. Stadler T: **Sampling-through-time in birth-death trees.** *J Theor Biol* 2010, **267:**396–404.
15. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, Günthard HF, Drummond AJ, Bonhoeffer S, Swiss HIV Cohort Study: **Estimating the basic reproductive number from viral sequence data.** *Mol Biol Evol* 2012, **29:**347–357.
16. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC: **Unifying the epidemiological and evolutionary dynamics of pathogens.** *Science* 2004, **303:**327–332.
17. Rasmussen DA, Ratmann O, Koelle K: **Inference for nonlinear epidemiological models using genealogies and time series.** *PLoS Comput Biol* 2011, **7:**1002136.
18. Ypma RJ, van Ballegooijen WM, Wallinga J: **Relating phylogenetic trees to transmission trees of infectious disease outbreaks.** *Genetics* 2013, **195:**1055–1062.
19. Felsenstein J: *Inferring Phylogenies.* Sunderland, MA: Sinauer Associates; 2004.
20. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT: **Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus.** *Proc Biol Sci* 2008, **275:**887–895.
21. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *New Engl J Med* 2011, **364:**730–739.
22. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG: **Measurably evolving populations.** *Trends Ecol Evol* 2003, **18:**481–488.
23. Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A: **Time-dependent rates of molecular evolution.** *Mol Ecol* 2011, **20:**3087–3101.
24. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of molecular evolution.** *Mol Biol Evol* 1998, **15:**1647–1657.
25. Yoder AD, Yang Z: **Estimation of primate speciation dates using local molecular clocks.** *Mol Biol Evol* 2000, **17:**1081–1090.
26. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.
27. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012, **61:**539–542.

28. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus.** *Nature* 2008, **453:**615–619.
29. Volz EM, Ionides E, Romero-Severson EO, Brandt M-G, Mokotoff E, Koopman JS: **HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis.** *PLoS Med* 2013, **10:**1001568.
30. Rasmussen DA, Boni MF, Koelle K: **Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam.** *Mol Biol Evol* 2014, **31:**258–271.
31. Kühnert D, Stadler T, Vaughan TG, Drummond AJ: **Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model.** *J R Soc Interface* 2014, **11:**20131106.
32. Volz EM: **Complex population dynamics and the coalescent under neutrality.** *Genetics* 2012, **190:**187–201.
33. Frost SD, Volz EM: **Modelling tree shape and structure in viral phylodynamics.** *Philos Trans R Soc London B Biol Sci* 2013, **368:**20120208.
34. Rasmussen DA, Volz EM, Koelle K: **Phylodynamic inference for structured epidemiological models.** *PLoS Comput Biol* 2014, **10:**1003570.
35. Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SD: **Phylodynamics of infectious disease epidemics.** *Genetics* 2009, **183:**1421–1430.
36. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD: **Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection.** *PLoS Comput Biol* 2012, **8:**1002552.
37. Koelle K, Rasmussen DA: **Rates of coalescence for common epidemiological models at equilibrium.** *J R Soc Interface* 2012, **9:**997–1007.
38. Stadler T, Bonhoeffer S: **Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods.** *Philos Trans R Soc Lond B Biol Sci* 2013, **368:**20120198.
39. Stadler T, Yang Z: **Dating phylogenies with sequentially sampled tips.** *Syst Biol* 2013, **62:**674–688.
40. Stadler T, Ku¨hnert D, Bonhoeffer S, Drummond AJ: **Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV).** *Proc Natl Acad Sci U S A* 2013, **110:**228–233.
41. Magiorkinis G, Sypsa V, Magiorkinis E, Paraskevis D, Katsoulidou A, Belshaw R, Fraser C, Pybus OG, Hatzakis A: **Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics.** *PLoS Comput Biol* 2013, **9:**1002876.
42. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz W: **Superspreading and the effect of individual variation on disease emergence.** *Nature* 2005, **438:**355–359.
43. Wallinga J, Lipsitch M: **How generation intervals shape the relationship between growth rates and reproductive numbers.** *Proc R Soc Biol Sci* 2007, **274:**599–604.
44. Svensson Å: **A note on generation times in epidemic models.** *Math Biosci* 2007, **208:**300–311.
45. Kenah E, Lipsitch M, Robins JM: **Generation interval contraction and epidemic data analysis.** *Math Biosci* 2008, **213:**71–79.
46. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, Posada D, Peeters M, Pybus OG, Lemey P: **HIV epidemiology. The early spread and epidemic ignition of hiv-1 in human populations.** *Science* 2014, **346:**56–61.
47. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sá-Leão R, Song JH, Ko KS, Beall B, Klugman KP, Parkhill J, Tomasz A, Kristinsson KG, Bentley SD: **Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone.** *BMC Biol* 2014, **12:**49.
48. Bedford T, Cobey S, Beerli P, Pascual M: **Global migration dynamics underlie evolution and persistence of human influenza A (H3N2).** *PLoS Pathog* 2010, **6:**1000918.
49. Bahl J, Nelson MI, Chan KH, Chen R, Vijaykrishna D, Halpin RA, Stockwell TB, Lin X, Wentworth DE, Ghedin E, Guan Y, Peiris JS, Riley S, Rambaut A, Holmes EC, Smith GJ: **Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans.** *Proc Natl Acad Sci U S A* 2011, **108:**19359–19364.
50. Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SY, Shapiro B, Pybus OG, Allain J-P, Hatzakis A: **The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis.** *PLoS Med* 2009, **6:**1000198.
51. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns AM, Hill RL, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramírez S, Feil EJ, Hudson LO, Enright MC, Balloux F, *et al*: **A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic.** *Genome Res* 2013, **23:**653–664.
52. Didelot X, Gardy J, Colijn C: **Bayesian inference of infectious disease transmission from whole genome sequence data.** *Mol Biol Evol* 2014, **31:**1869–1879.
53. Grassly NC, Harvey PH, Holmes EC: **Population dynamics of hiv-1 inferred from gene sequences.** *Genetics* 1999, **151:**427–438.
54. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH: **The epidemic behavior of the hepatitis C virus.** *Science* 2001, **292:**2323–2325.
55. Strimmer K, Pybus OG: **Exploring the demographic history of DNA sequences using the generalized skyline plot.** *Mol Biol Evol* 2001, **18:**2298–2305.
56. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, *et al*: **Emergence of Zaire Ebola virus disease in Guinea–preliminary report.** *New Engl J Med* 2014, **371:**1418–1425.
57. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P-P, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, *et al*: **Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.** *Science* 2014, **345:**1369–1372.
58. Dudas G, Rambaut A: **Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak.** *PLoS Curr* 2014, **6:**.
59. Center for Disease Control and Prevention: **Middle East Respiratory Virus (MERS).** 2014, [http://www.cdc.gov/coronavirus/mers/]
60. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enou V, van der Werf S, Ferguson NM: **Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility.** *Lancet Infect Dis* 2014, **14:**50–56.
61. Memish ZA, Mishra N, Olival KJ, Fagbo SF, Kapoor V, Epstein JH, Alhakeem R, Durosinloun A, Al Asmari M, Islam A, Kapoor A, Briese T, Daszak P, Al Rabeeah AA, Lipkin WI: **Middle East respiratory syndrome coronavirus in bats, Saudi Arabia.** *Emerg Infect Dis* 2013, **19:**1819–1823.
62. Haagmans BL, Al Dhahiry SH, Reusken CB, Raj VS, Galiano M, Myers R, Godeke GJ, Jonges M, Farag E, Diab A, Ghobashy H, Alhajri F, Al-Thani M, Al-Marri SA, Al Romaihi HE, Al Khal A, Bermingham A, Osterhaus AD, AlHajri MM, Koopmans MP: **Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation.** *Lancet Infect Dis* 2014, **14:**140–145.
63. Azhar EI, Hashem AM, El-Kafrawy SA, Sohrab SS, Aburizaiza AS, Farraj SA, Hassan AM, Al-Saeed MS, Jamjoom GA, Madani TA: **Detection of the Middle East respiratory syndrome coronavirus genome in an air sample originating from a camel barn owned by an infected patient.** *Mbio* 2014, **5:**e01450–14.
64. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF: **Rooting the phylogenetic tree of Middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat.** *J Virol* 2014, **88:**11297–11303.
65. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DM, Checchi F, Garcia E, Hugonnet S, Roth C, WHO Rapid Pandemic Assessment Collaboration: **Pandemic potential of a strain of influenza A (H1N1): early findings.** *Science* 2009, **324:**1557–1561.
66. Hedge J, Lycett S, Rambaut A: **Real-time characterization of the molecular epidemiology of an influenza pandemic.** *Biol Lett* 2013, **9:**20130331.
67. Wilson DJ: **Insights from genomics into bacterial pathogen populations.** *PLoS Pathog* 2012, **8:**1002874.
68. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ: **Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study.** *Lancet Infect Dis* 2013, **13:**130–136.
69. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, Van Wijngaerden E, Vandamme A-M, Van Laethem K, Lemey P: **The**

genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLoS Comput Biol* 2014, **10:**1003505.
70. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E: **Equation of state calculations by fast computing machines.** *J Chem Phys* 2004, **21:**1087–1092.
71. Fisher RA: *The Genetical Theory of Natural Selection.* Oxford: Clarendon; 1930.
72. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16:**97–159.
73. Griffiths RC, Tavare S: **Sampling theory for neutral alleles in a varying environment.** *Phil Trans R Soc Lond Biol Sci* 1994, **344:**403–410.
74. Pybus OG, Rambaut A, Harvey PH: **An integrated framework for the inference of viral population history from reconstructed genealogies.** *Genetics* 2000, **155:**1429–1437.
75. Minin VN, Bloomquist EW, Suchard MA: **Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics.** *Mol Biol Evol* 2008, **25:**1459–1471.

doi:10.1186/s13059-014-0541-9
Cite this article as: Li *et al.*: Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biology* 2014 **15**:541.