# Pathogen Epidemiology

**WP Hanage,** Harvard T.H. Chan School of Public Health, Boston, MA, USA

## Glossary

**Basic reproductive number** Often written $R_0$, this is the average number of onward transmission events resulting from the introduction of a single index case into a completely susceptible population.

**Coalescent** The Coalescent is a retrospective population genetic model that attempts to trace all alleles of a gene in a sample population to their most recent common ancestor (MRCA). This produces a gene genealogy. Coalescent theory seeks to understand the statistical properties of this genealogy under different selective or demographic scenarios.

**Incidence** New cases of disease occurring within a specified time period.

**Malthusian fitness** Where the population is growing in size, individuals that reproduce more quickly gain an advantage and will come to predominate even if their total net number of onward infections is less than more slowly growing competitors.

**Phylodynamics** The use of phylogenetic data to infer elements of epidemiological processes that have given rise to it, by examining the shape and topology of the genealogy as estimated from sequence data (see 'The Coalescent').

**Prevalence** The frequency of a disease in a population at a particular time point. Often expressed as a proportion or percentage.

**Reproductive number** Also known as the reproductive ratio or rate, this is the average number of onward transmission events produced from each infection.

**SIR model** In an SIR model hosts move from the susceptible compartment (S), to the Infected (I) and then (if they survive infection) into a separate resistant (R) compartment. This allows us to model immunity, by allowing hosts in the R compartment to be less likely to be successfully infected. An extra level of complexity can allow immunity to wane over time such that eventually hosts in the R compartment return to the S one – resulting in an SIRS model.

**SIS model** In an SIS model, hosts initially susceptible (S) become infected (I) and upon clearing the infection immediately becomes susceptible once more (S). Such simple models may be used where little or no immunity results from infection, or may be a useful approximation in the case that antigenic variation on the part of the pathogen is so rapid that immunity is negligible.

**Superspreader** An individual who infects a disproportionately large number of secondary hosts relative to the majority of cases of infection.

**Zoonoses** Any case where transmission of infectious agents occurs between different species of animals, most often used when disease is transmitted from nonhumans to humans.

Modern infectious disease epidemiology contains two distinct intellectual lineages: mathematical models are used to explore disease dynamics and the consequences of interventions (Anderson and May, 1991; Grassly and Fraser, 2008), while analysis of the genetic variation in pathogen populations is used to 'type' and define strains associated with resistance, virulence, or other features of interest. Recently these have been starting to come together with population genetic analyses of genetic (and increasingly genomic) variation that can tell us about the history of the sampled sequences. This has been termed 'phylodynamics' as a result of the combination of phylogenetic methods with the study of disease dynamics (Grenfell *et al.*, 2004).

Epidemiology is a population science that studies the patterns of disease incidence, attempting to infer causes and consequences. Classical epidemiology, for instance, might seek to identify risk factors for a given condition, which might be environmental or genetic. This allows us to identify interventions to minimize the risk of disease. In the case of a transmissible disease, the situation is somewhat different. The spread of infectious disease is a dynamic process in which the increasing numbers of cases increase the risk to the rest of the population. Similarly, as people recover they may become immune (or more pessimistically they may succumb to disease – it makes little difference to the infecting organism) and be removed from the system. Hence the numbers of hosts available to be infected change over time. To describe the changing state of the population, infectious disease epidemiology makes use of mathematical models comprising sets of differential equations, or statistical models that are defined using a probabilistic framework. These models can then be used to explore the impact of vaccination or other interventions.

## The Reproductive Number

A crucial parameter in infectious disease epidemiology is *R*: how many successful transmission events and new infections result on average from one infection. With an unlimited pool of susceptible individuals, this is equivalent to $R_0$ or the basic reproductive number. It is easy to intuitively relate these numbers to the course of an outbreak. If the reproductive number is $>1$, the expected number of new cases will increase, whereas if it is $<1$, then the numbers will fall. It is important to distinguish the incidence from the prevalence. Incidence is the number of new cases per unit time, whereas prevalence is the overall frequency of the disease in the population. The incidence can be falling but the prevalence can continue to

doi:10.1016/B978-0-12-800049-6.00228-6

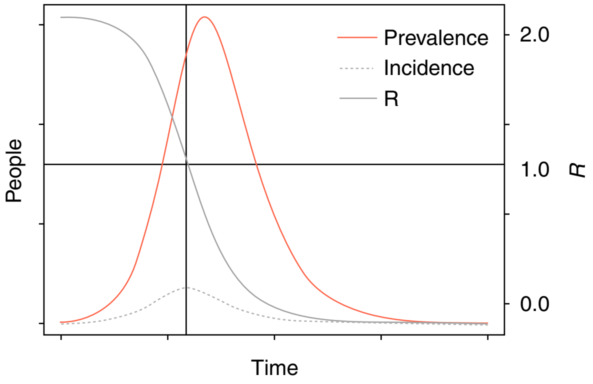

**Figure 1** An idealized epidemic curve showing how the prevalence, incidence, and reproductive number *R* vary over the course of an outbreak. The illustrated case is the result of an SIR model, in which recovered hosts become entirely resistant to infection and as a result the prevalence returns to zero. The point at which the decline in the availability of susceptible hosts means each case causes on average just one onward infection is that where $R=1$, and is indicated. Note that this is coincident with the peak incidence, which precedes peak prevalence as described in the text.

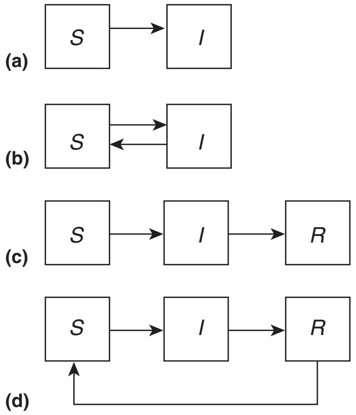

**Figure 2** Four simple compartmental model structures referred to in the text: (a) shows an SI model in which hosts become infected and never recover; (b) shows an SIS model in which recovery is possible as shown by the additional arrow; (c) shows an SIR model in which a recovered population is resistant to infection; and (d) a SIRS model incorporating waning immunity. The rates with which hosts transition between compartments can then be described using differential equations.

rise, albeit at a slower rate. This is illustrated in Figure 1, which shows an idealized epidemic curve, and how *R* changes over the course of the outbreak. While in the illustrated case the epidemic burns out after the pathogen has run out of hosts to infect, if sufficient susceptible hosts are continually introduced (e.g., by birth or waning immunity) the disease can become endemic.

Although the expected final size of the outbreak falls rapidly as $R_0$ decreases, it is possible for outbreaks to occur in the case where $R_0$ is less than 1. To understand how this can happen we have to remember that $R_0$ is an average value, and by chance initial cases may result in an above-average number of onward infections. The effects of this can be probed using a stochastic approach, in which events are modeled as randomly sampled realizations from a probability distribution. While such models aim to capture the typical contact and transmission dynamics in a population, they can be confounded by rare cases of extreme behavior. For instance, a so-called 'superspreader' event may by chance infect many new hosts, starting an outbreak that persists until all the transmission chains descending from it die out (Lloyd-Smith *et al.*, 2005; Garske and Rhodes, 2008). For a real example of such a superspreader event, consider the early stages in the 2003 SARS (severe acute respiratory syndrome) outbreak in Hong Kong, when a single infected person infected at least 13 and possibly as many as 20 others staying at the same hotel (Braden *et al.*, 2013). The estimated $R_0$ for SARS is much lower than this unusual event would lead you to suspect (Riley *et al.*, 2003).

We can construct a simple transmission model by dividing the population into compartments, and then writing down equations for the rate of movement of individuals from one compartment to another in each time step. The simplest models contain just two compartments: susceptible and infected (Figure 2(a) and 2(b)). The results are a so-called SI or SIS model. In the first example, hosts move from being susceptible to being infected, and in the second, to infected hosts can clear the infection and become susceptible again. The slightly more complicated SIR model adds a 'resistant' compartment (Figure 2(c)), reflecting immunity that prevents infection while waning immunity can be incorporated in a SIRS framework where after a period of time hosts in the resistant compartment move back to S (Figure 2(d)). This flexible approach can be extended to include important features such as vaccination (McLean, 1995) (in the simplest case of a 100% effective vaccine, each vaccinated susceptible would move to the resistant compartment), and host population structure. This is important in multiple contexts; for instance the child–child contact rate is very high in daycare settings and hence the transmission rate too, which can have important consequences for model results (Schenzle, 1984). Similarly, humans vary greatly in the number and sort of sexual contacts they make, which is important to modeling sexually transmitted infections. In the case of vector-borne diseases, whether the vector is a biting insect or a health care worker (as in nosocomial infections), they can and should be incorporated into the model. It is easy to see intuitively how the endemic prevalence of disease depends on the supply of new susceptibles, either from birth or waning immunity.

## Transmission Routes and the Target of Selection

Infectious agents can be categorized by how they get from one host to another, and whether this involves any intermediate hosts or environments. Some pathogens spread without spending significant time in the environment, examples being influenza or sexually transmitted infections. Vector-borne diseases in contrast rely on an additional host for transmission, often a biting insect. A considerable burden of disease also results from pathogens that transmit between humans rarely if at all. These infections arise from environmental exposure to the pathogen, and can include zoonoses – infections acquired from other species. In each case it is useful to consider which traits of the pathogen will be scrutinized by selection.

In the case of directly transmitted pathogens it is of paramount importance to colonize new hosts; in other words, maximize the reproductive number. Note that this does not necessarily mean causing disease, merely transmission, although some features of disease such as sneezing may be adaptive traits. As immunity builds in the population the effective reproductive number will fall, and one possible means of increasing it is immune escape, and there is typically evidence of strong diversifying selection on those antigens that can yield protective immunity (Li *et al.*, 1995). Vector-borne pathogens must successfully survive in and transmit between the host and the vector species, and *vice versa*. Again, we expect and observe strong diversifying selection on those antigens targeted by the immune system.

One area of particular interest is the evolution of virulence. It is commonly stated that virulence reflects an adaptive mismatch between the host and the pathogen: it is to the advantage of both parties that virulence be minimal, such that the host can continue to transmit for a long period of time without limits on the numbers of contacts they make. Empirical evidence for this comes from the example of myxomatosis in Australia, where the myxoma virus was introduced to control rabbit populations. With extraordinary vision, Frank Fenner of the Australia National University collected and stored isolates of virus collected from wild rabbits over decades, and showed evidence for both gradual attenuation of the virus, and adaptation of the rabbit population to become more resistant (Fenner, 1956).

While this argument sounds persuasive, it is also regarded with suspicion as an example of group selection. Surely more rapidly growing parasite variants should be selected during infection and, if this is linked to virulence, produce disease (Levin and Bull, 1994)? The tension between the two selective pressures – for growth within the host, and for transmission within the host population – is dependent on the link between virulence and transmission (Lipsitch and Moxon, 1997; Vale *et al.*, 2011). If by causing disease, a pathogen transmits to more new hosts than it would do otherwise, then virulence will be increased even if the result is the death of the host or the clearance of the infection (Anderson and May, 1982). This explains virulence in terms of the theory of life-history tradeoffs. A crucial factor is the availability of new hosts, because if opportunities for transmission are rare it will select for variants that do not kill their hosts before they can transmit (Lipsitch and Nowak, 1995; Lenski and May, 1994). In an outbreak situation, virulence can be temporarily selected even if the virulent strain infects fewer hosts per infection than its competitors, provided it does so more quickly. Selection for more rapidly growing variants is also termed Malthusian fitness (Orr, 2009).

The question of whether virulence is adaptive is not of merely academic interest. HIV-1 infection includes a long asymptomatic period in which the viral load (the amount of virus particles in peripheral blood) fluctuates around a value known as the 'set point' (Mellors *et al.*, 1996). It has been proposed that the set point viral load has been the object of selection to maximize transmission, the result of a trade off between the effect of set point viral load on infectiousness and the duration of the asymptomatic period (Fraser *et al.*, 2007). If set point viral load is too high, then infected individuals have fewer opportunities to transmit before developing AIDS whereas if it is too low, they are not infectious enough to efficiently transmit the virus. For natural selection to act on set point viral load, it must be a heritable property. That is to say, the set point viral load of the infector and infected must be correlated. If it is not, then there is no heritable variation on which natural selection can work. Subsequent work has produced evidence that set point viral load is indeed heritable, and correlated with viral genotype (Hollingsworth *et al.*, 2010; Alizon *et al.*, 2010).

Virulence has also been suggested to be adaptive in malaria transmission. Here, the greater the numbers of parasites in peripheral blood (higher parasitemia), the greater the chance that a mosquito becomes infected while feeding. Parasitemia is also correlated with severity of disease. If disease can limit onward transmission, 'imperfect' or partially effective vaccines have been proposed to select for strains capable of growing to a higher titer in both hosts receiving the therapy and those who are not (Gandon *et al.*, 2001; Barclay *et al.*, 2012). This could lead to an increase in virulence in the absence of vaccination – and given known inequities in access to health care, such a perverse outcome could have a major impact on some patient populations. However, unpicking which factors contribute to virulence is extremely difficult and the goal of virulence management remains an aspiration rather than an actuality (Dieckmann and International Institute for Applied Systems Analysis, 2005).

## Molecular Epidemiology: The Short and Long Term

It is often important to be able to link individual cases of disease on the basis of the similarity of the infecting pathogen, such as in defining an outbreak or identifying drug resistant lineages. These are examples respectively of short-term and long-term epidemiology. To address such questions, the variation in the pathogen population must be assayed in some way, and the results compared. This is now almost always done using molecular variation, i.e., nucleic acid sequences and the proteins they encode. How much variation is present in the population, and of what sort, is the result of evolution, and how data from molecular epidemiology studies are interpreted is a question of phylogenetics. Increasingly, populations of viruses and bacteria are characterized using genomic methods, and it is likely that in the near future whole genomes will become the standard for molecular epidemiology (Croucher *et al.*, 2013).

Molecular epidemiology catalogs the variation in the pathogen population as different 'types' that can be distinguished. Historically these have often been defined using antibodies to distinguish between different variants of cell surface markers and divide the population into serotypes. Pathogens of all kinds have been distinguished using this approach: viruses, bacteria, and protozoa. When we speak of a case of influenza caused by 'H5N1,' we are referring to the serotypes of the surface proteins hemagglutinin and neuraminidase, which in this case are those associated with 'avian flu.' In *Escherichia coli* and *Salmonella*, the 'O' antigens (oligosaccharides) are combined with the 'H' antigens (flagellar proteins) to provide a discriminating serotyping scheme.

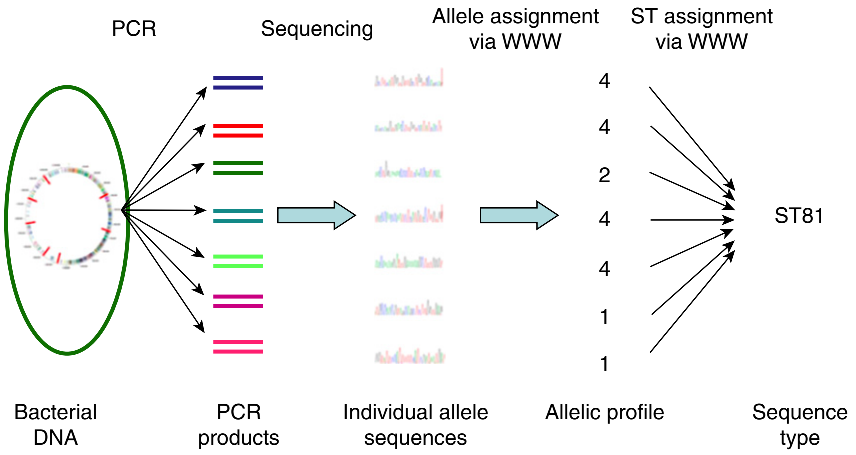

**Figure 3** Illustration of the Multi-Locus Sequence Typing approach. The sequences of multiple housekeeping loci are determined, and then allele numbers assigned via an internet database, which together make up the allelic profile. This in turn determines the Sequence Type or ST, also assigned by interrogating an internet database. The example shown is for ST 81 in the *Streptococcus pneumoniae* typing scheme. While the illustration shows sequences determined by Sanger sequencing, increasingly data from genomic approaches are accepted.

*Salmonella enterica* alone can be divided into more than 2500 serovars (Grimont and Weill, 2007). The O:1 and O:139 serotypes of *Vibrio cholerae* produce a phage-borne toxin, cholera toxin, which causes the disease that bears its name (Finkelstein, 1996). More than 200 serotypes of *V. cholerae* are known that do not produce cholera toxin or cause cholera (Shimada *et al.*, 1994).

Serotyping and other typing methods that assay phenotypes suffer from limited discrimination: they use a tiny fraction of the diversity associated with a strain, and the variation they assay is frequently produced by intense diversifying selection from the immune system. The ideal data for typing are unambiguous and easily portable between labs. We also want to be able to describe how the types we identify are related. We can group serologically related strains as serovars or serogroups, but the relationships within and between these cannot be ascertained in detail. A single serological type can contain many different genotypes, and any horizontal transfer of the locus determining serotype can lead to distantly related lineages being indistinguishable by this method.

Outbreak analysis is a short-term question requiring highly discriminating methods. The response is to assay multiple, rapidly changing regions in the entire genome. Examples of such regions are restriction sites, sites at which PCR primers can bind, or regions of repeat sequences, where the numbers of repeats can change rapidly. Pulsed Field Gel Electrophoresis (PFGE), which uses restriction sites, is an excellent example. Genotypes are distinguished by banding patterns on a gel resulting from changes in the position and numbers of restriction sites. PFGE remains a commonly used method in this context (e.g., Choi *et al.*, 2014).

PFGE and related methods are less useful for evolutionary or population genetics. The selective impacts of the changes are unknown, because we do not know where the restriction sites lie in the genome. While we can identify closely related banding patterns, beyond very closely related strains relationships become hard or impossible to discern. Practically it is difficult to compare results between laboratories. As sequencing has become easier and more accessible, it has become standard to use nucleic acids as the source of assayed variation. In bacteria, a popular approach is to sequence multiple loci scattered around the genome that encode core metabolic or 'housekeeping' functions. This allows the analysis of synonymous SNPs that are unlikely subjects for diversifying selection. The approach, first applied to *Neisseria meningitidis*, is termed multi-locus sequence typing (MLST) (Maiden *et al.*, 1998), and is illustrated in **Figure 3**.

Unlike bands on gels, the sequence data used by MLST are unambiguous. This means that the effort of collecting data on the allelic variants found in the community can be distributed to researchers worldwide. Individual labs in remote locations can determine the sequence of the loci used in the MLST scheme, and then compare them with an online database (mlst.net and pubmlst.net). If the allele is novel, it may be added to the database for future users. Each isolate is defined by the combination of alleles at the MLST loci. Each unique allele is identified by an integer, and the combination of these makes up the allelic profile and the sequence type (ST). MLST has been applied to numerous organisms. One of the side effects of the data collected by epidemiologists has been in the study of homologous recombination (reviewed in this volume by Feil). The wealth of discriminating data collected for MLST has shown that in many named species, recombination is a more frequent source of change at the MLST loci than mutation.

## Genomic Epidemiology

The genomic revolution promises to have a profound effect on molecular epidemiology. While in the past, whole genome analyses were the preserve of virologists, it is now economical and, more importantly, easy to obtain quality data on a far higher proportion of the bacterial (or protozoal) genome than afforded by previous methods. Genomic methods have been applied to outbreaks of diseases including tuberculosis (Gardy *et al.*, 2011; Walker *et al.*, 2013), *E. coli* infections (Grad *et al.*, 2012; Mellmann *et al.*, 2011) and cholera (Katz *et al.*, 2013;

Eppinger *et al.*, 2014), and are becoming folded into the routine epidemiologic work of public health authorities.

The pace of change in this field is such that any detailed discussion of technology will be shortly superseded. Nevertheless, important principles can be defined. At present multiple technologies for sequencing exist, and moreover, multiple approaches for putting raw data together to make a genome. The great majority of studies are more properly called 'genomic' rather than 'whole genome' methods, because in hardly any cases have genomes been completely sequenced. Instead, a very high proportion of the genome is determined. While high-quality draft genomes can be used for many interesting things, they are not finished: we do not know the sequence of the chromosome all the way round from the origin of replication.

Which parts of the genome get missed? This depends on the sequencing platform and the methods of assembly, but typically highly unstable regions such as repeats are problematic. These are of course the regions that are most useful for short-term questions in epidemiology. In their absence, we can look at single nucleotide polymorphisms elsewhere in the genome. In comparing very closely related isolates that may differ at a handful of SNPs, the possibility of false positives becomes acute, so we must deal with the fact that different technologies and analytic approaches have different error rates (Croucher *et al.*, 2013).

Taken together, the profusion of genomic methods means that rather than removing ambiguity in the comparison of closely related isolates, new sources of ambiguity have been discovered. At the level of resolving more distantly related isolates into major lineages, equivalent to MLST, genomics has been highly successful and revealed both considerable variation within closely related STs, and perhaps surprisingly shown that MLST in most cases effectively identified the major lineages. The potential of methods that can sequence through unstable repeat regions with high fidelity is real, but has not been conclusively shown at the time of writing. This is likely to change, but for the benefit to be felt the technology will have to be cheap enough for many labs to use.

## Phylodynamics and Using DNA Sequence to Study Transmission

The most exciting recent development at the interface of evolutionary biology and epidemiology has come about from the proliferation of sequence data combined with methods capable of making inferences about the history of a sample of sequences from the structure of the genealogy underlying them. A key concept is the coalescent (Kingman, 1982), which describes the genealogy of a sample of sequences in terms of how often their lineages 'coalesce' or come together to form an internal node in the tree. Combined with a molecular clock that relates the accumulation of sequence divergence to time, this allows us to infer events that have happened in the history of the sequences, most notably and relevant for epidemiology, changes in population size. The basic rationale is readily grasped, and accessible software is available to implement the methods (Drummond *et al.*, 2012; Bouckaert *et al.*, 2014).

As stated above we describe the structure of a genealogy by looking at how often the lineages coalesce, relative to the branch length, which in the case of a molecular clock is a proxy for time. The most valuable data are sequences sampled over time, providing 'measurably evolving populations' that can be used to estimate the clock rate (Ewing *et al.*, 2004). An alternative approach is to model transmission as a birth death process, as discussed in Boskova *et al.* (2014).

The rate of coalescence over the tree is simply related to population size by reflecting that in a smaller population individuals are more likely to share a parent in the previous generation by chance. In fact the probability two individuals share the same parent in the previous generation is the reciprocal of population size. As a result the rate of coalescence is higher in smaller populations. This allows inference as to whether the population is expanding, has experienced a bottleneck, or any of multiple other demographic and other processes.

The ability to study changing population size from sequence alone can be used, with an estimate of the serial interval or time between infection and transmission, to estimate the reproductive number. Analysis of sequence variation has been used to study pathogens including Influenza (Hedge *et al.*, 2013), *V. cholerae* (Katz *et al.*, 2013), MERS (Cauchemez *et al.*, 2014) and the recent Ebola outbreak (Gire *et al.*, 2014). This work has shown the potential of the approach. While these analyses are readily approached using the BEAST suite of programs (Bouckaert *et al.*, 2014; Drummond *et al.*, 2012), as usual it should not be assumed that default parameters are appropriate.

Using sequence data to infer demographic history is distinct from using it to infer recent transmission. At the extremely local level it is hoped that we might be able to define infector and infected using high-resolution sequence data alone. Indeed, highly resolving methods that stop short of the whole genome have been used as a valuable complement to classical contact tracing for the control of gonococcal disease (Bilek *et al.*, 2007). At a higher level we might ask whether cases of disease occurring some distance apart are due to closely related pathogens, which might imply transmission over greater distances – for an example, involving wind-borne transmission of avian influenza see Ypma *et al.* (2013b). The overall principle is that the transmission tree is considered closely related to the phylogenetic tree (Ypma *et al.*, 2013c). To be useful this need not be the precise resolution of who infected whom; different epidemiological events like superspreading (Ypma *et al.*, 2013a) may leave distinct signatures in the resulting tree structure (Colijn and Gardy, 2014). Software packages are becoming available that implement multiple methods to probe sequence data to provide the detailed history of an outbreak (Jombart *et al.*, 2014a,b). However, concerns exist over the potential for within-host evolution to produce sequence diversity that can obscure true transmission networks (Didelot *et al.*, 2014; Worby *et al.*, 2014). The amount of diversification that occurs within the host is not known in most cases, but may be large (Nasser *et al.*, 2014). Coupled with the uncertainty arising from imperfect sampling and other missing data, it is likely that the most valuable uses of sequence data will be as a complement to traditional epidemiologic approaches (Ypma *et al.*, 2012).

## Conclusion

Infectious disease epidemiology is a practical science, concerned with minimizing the impact of pathogens on public health. As both pathogens and their hosts have evolved, evolutionary biology is relevant to understanding the nature of their interactions for fitness, and also in resolving the history of pathogen transmission. Mathematical models can explore the consequences of different selective scenarios, and molecular data can define strains and the genetic variation that is the raw material on which natural selection acts. Recent advances, especially in the rapid determination of sequence data, are bringing evolutionary biology ever closer to the clinic.

*See also*: Evolutionary Medicine IV. Evolution and Emergence of Novel Pathogens. Recombination in Bacterial Populations

## References

Alizon, S., Von Wyl, V., Stadler, T., *et al.*, 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. PLoS Pathogens 6, e1001123.

Anderson, R.M., May, R.M., 1982. Coevolution of hosts and parasites. Parasitology 85 (Pt 2), 411–426.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford; New York, NY: Oxford University Press.

Barclay, V.C., Sim, D., Chan, B.H., *et al.*, 2012. The evolutionary consequences of blood-stage vaccination on the rodent malaria *Plasmodium chabaudi*. PLoS Biology 10, e1001368.

Bilek, N., Martin, I.M., Bell, G., *et al.*, 2007. Concordance between *Neisseria gonorrhoeae* genotypes recovered from known sexual contacts. Journal of Clinical Microbiology 45, 3564–3567.

Boskova, V., Bonhoeffer, S., Stadler, T., 2014. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. PLoS Computational Biology 10, e1003913.

Bouckaert, R., Heled, J., Kuhnert, D., *et al.*, 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10, e1003537.

Braden, C.R., Dowell, S.F., Jernigan, D.B., Hughes, J.M., 2013. Progress in global surveillance and response capacity 10 years after severe acute respiratory syndrome. Emerging Infectious Diseases 19, 864–869.

Cauchemez, S., Fraser, C., Van Kerkhove, M.D., *et al.*, 2014. Middle East respiratory syndrome coronavirus: Quantification of the extent of the epidemic, surveillance biases, and transmissibility. Lancet Infectious Diseases 14, 50–56.

Choi, M.J., Jackson, K.A., Medus, C., *et al.*, 2014. Notes from the field: Multistate outbreak of listeriosis linked to soft-ripened cheese–United States, 2013. Morbidity and Mortality Weekly Report 63, 294–295.

Colijn, C., Gardy, J., 2014. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health 2014, 96–108.

Croucher, N.J., Harris, S.R., Grad, Y.H., Hanage, W.P., 2013. Bacterial genomes in epidemiology–present and future. Philosophical Transactions of the Royal Society B: Biological Sciences 368, 20120202.

Didelot, X., Gardy, J., Colijn, C., 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. Molecular Biology and Evolution 31, 1869–1879.

Dieckmann, U., International Institute for Applied Systems Analysis, 2005. Adaptive Dynamics of Infectious Diseases: In Pursuit of Virulence Management. New York, NY: Cambridge University Press.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29, 1969–1973.

Eppinger, M., Pearson, T., Koenig, S.S., *et al.*, 2014. Genomic epidemiology of the haitian cholera outbreak: A single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. MBio 5, e01721.

Ewing, G., Nicholls, G., Rodrigo, A., 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 168, 2407–2420.

Fenner, F., 1956. Evolutionary aspects of Myxomatosis in Australia. Memórias do Instituto Oswaldo Cruz 54, 271–278.

Finkelstein, R.A., 1996. Cholera, *Vibrio cholerae* O1 and O139, and other pathogenic vibrios. In: Baron, S. (Ed.), Medical Microbiology, fourth ed. Galveston, TX: University of Texas Medical Branch at Galveston, pp. 14–27.

Fraser, C., Hollingsworth, T.D., Chapman, R., De Wolf, F., Hanage, W.P., 2007. Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. Proceedings of the National Academy of Sciences of the United States of America 104, 17441–17446.

Gandon, S., Mackinnon, M.J., Nee, S., Read, A.F., 2001. Imperfect vaccines and the evolution of pathogen virulence. Nature 414, 751–756.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., *et al.*, 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine 364, 730–739.

Garske, T., Rhodes, C.J., 2008. The effect of superspreading on epidemic outbreak size distributions. Journal of the Theoretical Biology 253, 228–237.

Gire, S.K., Goba, A., Andersen, K.G., *et al.*, 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345, 1369–1372.

Grad, Y.H., Lipsitch, M., Feldgarden, M., *et al.*, 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proceedings of the National Academy of Sciences of the United States of America 109, 3065–3070.

Grassly, N.C., Fraser, C., 2008. Mathematical models of infectious disease transmission. Nature Reviews Microbiology 6, 477–487.

Grenfell, B.T., Pybus, O.G., Gog, J.R., *et al.*, 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303, 327–332.

Grimont, P.A.D., Weill, F.-X., 2007. Antigenic Formulae of the *Salmonella serovars*, Institut Pasteur. Paris, France: World Health Organization Collaborating Centre for Reference and Research on Salmonella.

Hedge, J., Lycett, S.J., Rambaut, A., 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biology Letters 9, 20130331.

Hollingsworth, T.D., Laeyendecker, O., Shirreff, G., *et al.*, 2010. HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. PLoS Pathogens 6, e1000876.

Jombart, T., Aanensen, D.M., Baguelin, M., *et al.*, 2014a. OutbreakTools: A new platform for disease outbreak analysis using the R software. Epidemics 7, 28–34.

Jombart, T., Cori, A., Didelot, X., *et al.*, 2014b. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Computational Biology 10, e1003457.

Katz, L.S., Petkau, A., Beaulaurier, J., *et al.*, 2013. Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti. MBio 4, e00398-13.

Kingman, J.F.C., 1982. On the genealogy of large populations. Journal of Applied Probability 19A, 27–43.

Lenski, R.E., May, R.M., 1994. The evolution of virulence in parasites and pathogens: Reconciliation between two competing hypotheses. Journal of the Theoretical Biology 169, 253–265.

Levin, B.R., Bull, J.J., 1994. Short-sighted evolution and the virulence of pathogenic microorganisms. Trends in Microbiology 2, 76–81.

Li, J., Ochman, H., Groisman, E.A., *et al.*, 1995. Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. Proceedings of the National Academy of Sciences of the United States of America 92, 7252–7256.

Lipsitch, M., Moxon, E.R., 1997. Virulence and transmissibility of pathogens: What is the relationship? Trends in Microbiology 5, 31–37.

Lipsitch, M., Nowak, M.A., 1995. The evolution of virulence in sexually transmitted HIV/AIDS. Journal of the Theoretical Biology 174, 427–440.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438, 355–359.

Maiden, M.C., Bygraves, J.A., Feil, E., *et al.*, 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America 95, 3140–3145.

McLean, A.R., 1995. Vaccination, evolution and changes in the efficacy of vaccines: A theoretical framework. Proceedings: Biological Sciences 261, 389–393.

Mellmann, A., Harmsen, D., Cummings, C.A., *et al.*, 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6, e22751.

Mellors, J.W., Rinaldo Jr., C.R., Gupta, P., *et al.*, 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science 272, 1167–1170.

Nasser, W., Beres, S.B., Olsen, R.J., *et al.*, 2014. Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. Proceedings of the National Academy of Sciences of the United States of America 111, E1768–E1776.

Orr, H.A., 2009. Fitness and its role in evolutionary genetics. Nature Reviews Genetics 10, 531–539.

Riley, S., Fraser, C., Donnelly, C.A., *et al.*, 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. Science 300, 1961–1966.

Schenzle, D., 1984. An age-structured model of pre- and post-vaccination measles transmission. IMA Journal of Mathematics Applied in Medicine and Biology 1, 169–191.

Shimada, T., Arakawa, E., Itoh, K., *et al.*, 1994. Extended Serotyping Scheme for *Vibrio cholerae*. Current Microbiology 28, 175–178.

Vale, P.F., Wilson, A.J., Best, A., Boots, M., Little, T.J., 2011. Epidemiological, evolutionary, and coevolutionary implications of context-dependent parasitism. American Naturalist 177, 510–521.

Walker, T.M., Ip, C.L., Harrell, R.H., *et al.*, 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. Lancet Infectious Diseases 13, 137–146.

Worby, C.J., Lipsitch, M., Hanage, W.P., 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Computational Biology 10, e1003549.

Ypma, R.J., Altes, H.K., Van Soolingen, D., Wallinga, J., Van Ballegooijen, W.M., 2013a. A sign of superspreading in tuberculosis: Highly skewed distribution of genotypic cluster sizes. Epidemiology 24, 395–400.

Ypma, R.J., Bataille, A.M., Stegeman, A., *et al.*, 2012. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings: Biological Sciences 279, 444–450.

Ypma, R.J., Jonges, M., Bataille, A., *et al.*, 2013b. Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. Journal of Infectious Diseases 207, 730–735.

Ypma, R.J., Van Ballegooijen, W.M., Wallinga, J., 2013c. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195, 1055–1062.