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
 http://dx.doi.org/10.1016/j.tim.2014.02.011

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.