Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

The application of genomics to tracing bacterial pathogen transmission

Nicholas J Croucher and Xavier Didelot

New sequencing technologies have made it possible to generate bacterial genomes at clinically relevant timescales and price levels. The use of whole-genome sequencing (WGS) has proved useful for investigating transmission at different scales. WGS data are highly effective at determining whether individuals are part of the same transmission chain, making it possible to detect probable direct transmission events, delimit the extent of local nosocomial or community-based outbreaks, and identify worldwide patterns of spread and long-term dynamics of bacterial pathogens. Making the most of WGS data will probably always require associated detailed epidemiological data, but nevertheless it promises to become an increasingly valuable tool for infection control in the near future.

Addresses
Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, United Kingdom

Corresponding author: Didelot, Xavier (x.didelot@imperial.ac.uk)

Current Opinion in Microbiology 2015, **23**:62–67

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay CD Hinton**

For a complete overview see the Issue and the Editorial

Available online 22nd November 2014

http://dx.doi.org/10.1016/j.mib.2014.11.004

Introduction

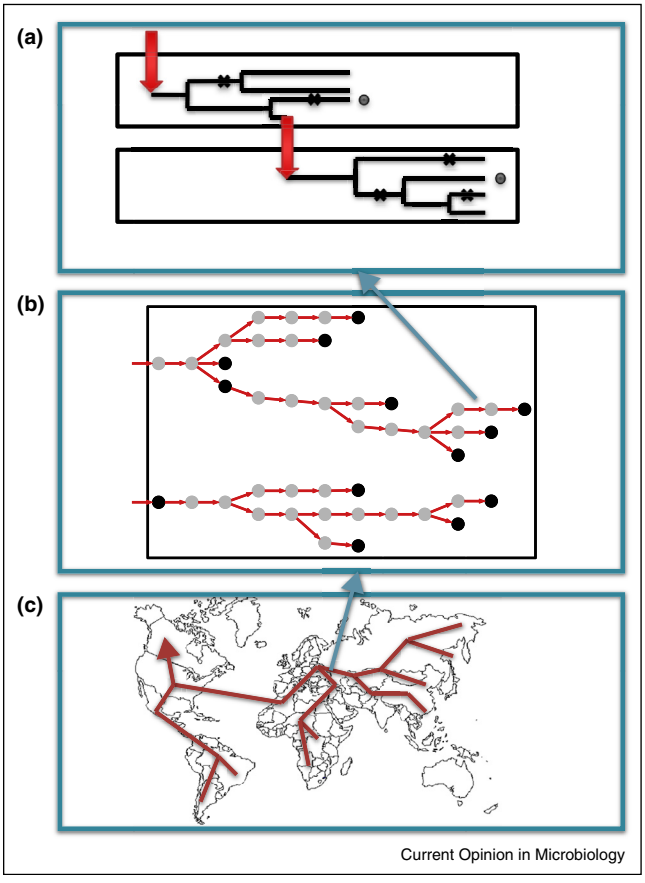
New sequencing technologies have made it possible to generate draft bacterial genomes within a day or two of sampling an isolate at clinically-relevant prices [1–3]. The greater resolution of whole genome sequencing (WGS) data relative to older genotyping methods makes it a potentially powerful method in infectious disease epidemiology [4]. WGS also has the practical advantage over older typing methods of being universally applicable to many bacterial pathogens, as demonstrated by many studies over recent years. These can be subdivided into three types, according to the scale at which transmission is studied: either directly between hosts, or indirectly in a single location, or globally (Figure 1).

Direct transmission between individuals

The finest scale of epidemiological investigation is to ask whether direct transmission happened from a given host to another. Dense sampling of bacterial isolates from a specific host population has been used to identify likely cases of direct transmission. Typically WGS is applied to only one isolate per host, and these sequences are analyzed in a pairwise fashion (Figure 1a). The overall distribution of pairwise distances has often proved to be multimodal [5••,6], with one peak representing the short distances between closely-related isolates that represent candidate transmission events. This allows a threshold distance to be defined beyond which direct transmission can be ruled out. For example, dense sampling of *Clostridium difficile* disease within the hospitals of a single UK county has been used to investigate nosocomial transmission [7•]. The transmission events likely to have occurred directly between symptomatic cases were detected based on isolate pairs that were distinguished by two or fewer single nucleotide polymorphisms (SNPs). However, there were many cases were no likely donor was found, suggesting that the majority of *C. difficile* transmission occurred via other routes, for example involving asymptomatic carriers. Similarly, studies of community outbreaks of the respiratory pathogen *Mycobacterium tuberculosis* have used a conservative threshold of 12 SNPs to rule out some suspected transmission events, which led to the identification of co-circulating, independent transmission chains [6,8]. Conversely, such studies have also identified previously unsuspected 'cryptic' transmission events [6,8,9], as well as 'superspreaders' that have transmitted to multiple other individuals [8]. WGS promises to be effective in detecting such persons [10], who represent important targets for infection control measures.

Rather than using the original genetic distances to infer transmission between isolate pairs, their divergence may instead be calculated as a time to their most common recent ancestor (TMRCA) using a molecular clock. This allows for biological considerations to be integrated into analyses, such as the known incubation period of *C. difficile* when studying nosocomial transmission [11]. It also permits non-clocklike genetic divergence to be taken into account, as shown in an analysis of familial transmission of *Helicobacter pylori* where high and variable recombination rates were found [12]. Excluding such sequence exchanges made it possible to infer the rate of within-host diversification through point mutation using multiple pairs of samples from the same hosts, facilitating the estimation of who infected whom. TMRCAs also allow for epidemiological constraints to be incorporated into analysis, for example using information on when contact

Figure 1



Conceptual representation of the three scales at which transmission is envisaged. **(a)** Direct transmission scale. Transmission events are shown as red arrows, and the trees represent within-host evolution including mutation events (black crosses). WGS data is typically collected for only one isolate per host (gray dots) at the time when the infection is detected and possibly eradicated. **(b)** Local transmission scale. WGS data is only available for a subset of infected hosts (black dots) that are interspersed within transmission chains composed mainly of unsampled hosts (gray dots). Multiple imports from external sources may have happened (two are shown here, represented by the leftmost arrows). **(c)** Global transmission scale.

was likely between individuals to help infer patterns of transmission [5••,13••].

Sequencing multiple colonies to measure the 'cloud' of bacterial diversity within a host [13••] can provide crucial information in determining transmission chains [14,15]. For example, seven isolates were sequenced from the first detected (index) case of a *Klebsiella pneumoniae* outbreak in a US clinical center, along with one representative from each of 18 other affected individuals [16]. This resolved the outbreak into three independent transmission chains, which associated with two different anatomical sites on the index case.

Phylogenetic analyses of multiple isolates from individuals can provide additional power to discriminate transmission chains, as putative transmissions can be ruled out by the topology of a phylogeny, even if they are closely related [17]. Transmission is most robustly inferred when the genetic diversity of one individual's isolates is derived from that of another individual (Figure 2). Such evidence of transmission has been observed in bacterial samples from cystic fibrosis (CF) patients, in whom longitudinal sampling is able to identify the comparatively high bacterial diversity that arises during chronic infection. This allowed robust evidence of direct transmission to be identified from a set of *Mycobacterium abscessus* sequences from CF patients in the UK, including the assignment of donors and recipients based on relative levels of diversity and associated epidemiological information [5••]. Analysis of an outbreak of *Burkholderia dolosa* among CF patients in Boston also identified such patterns in the reconstructed tree [18]; however, as not all individuals in the outbreak had been sampled, this could only be used to infer the relative ordering of individuals in the transmission chain, rather than providing strong evidence of direct transmission.

Genomics in local outbreak investigations

Typically only a small proportion of infected individuals are sampled in many transmission analyses conducted at the local scale of a hospital, town or country (Figure 1b). In such studies, there is always incomplete sampling of the infected hosts such that isolates are indirectly linked via a chain of unsampled hosts, and there may have been multiple imports of the pathogen into the local population [19].

Several local clusters of methicillin-resistant *Staphylococcus aureus* (MRSA) cases within healthcare systems have been studied using WGS [3,13••,20–22,23•]. Such data have been important in demonstrating that these sets of cases represent genuine outbreaks: in some instances, cryptic transmissions were uncovered [13••,20,22], and some individuals potentially considered part of the outbreak were found to be independent cases [20,21]. The advantages of WGS over previous methods are especially clear in examples where other information has resulted in doubt as to whether a single outbreak had occurred, as clusters of cases were sometimes separated by weeks and variation was observed between antibiograms of outbreak isolates [3,13••]. The intermittent nature of these outbreaks suggested a reservoir of infection within the hospital, and in two cases, screening of healthcare workers found individuals colonized with strains related to the outbreak [13••,21]. However, only when multiple colonies were sampled from the worker was it possible to infer their role in transmission [13••].

The potential of WGS for investigating local outbreaks from a point source was demonstrated during the 2011 outbreak of *Escherchia coli* O104:H4 caused by contaminated sprouts. Prospective sequencing of bacteria