Genomics update

Genomic tracing of epidemics and disease outbreaks

Anita C. Schürch1,2 and Roland J. Siezen2,3*
1RIVM, National Institute for Public Health and the Environment, 3730BA Bilthoven, the Netherlands.
2Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, 6500HB Nijmegen, the Netherlands.
3NIZO food research, 6710BA Ede, the Netherlands.

Tracing the source of an infectious human disease can save lives. It allows for measures to be taken to prevent further spread of the disease. Although the mode of transmission for many human pathogens is known, it often remains difficult to trace the exact source of an outbreak of a disease with laboratory methods. Viruses, bacteria, fungi, parasites and protozoa can cause human diseases, but here we focus on bacterial pathogens. The currently used techniques to obtain DNA fingerprints of bacterial agents of infectious diseases frequently cannot discriminate between all bacterial strains of the same outbreak, making it impossible to follow the spread of the disease. A recent solution to this problem is the application of next-generation whole-genome sequencing techniques, which allows all available genetic information of each clinical isolate to be determined.

Trends in bacterial typing

Historically, identification and classification of bacterial pathogens have been accomplished with phenotypic analyses, such as bacteriophage typing or drug susceptibility testing. Nowadays, molecular biology techniques such as restriction-fragment length polymorphism typing [RFLP (Todd et al., 2001)] or pulsed-field gel electrophoresis are used to assign a ‘type’ to a bacterial isolate, together with techniques that rely on variations in sequence repeat lengths [variable numbers of tandem repeats, VNTR (van Belkum, 1999)], or on sequencing of one or several housekeeping genes, for example spa typing (Frenay et al., 1996) or multilocus sequencing typing [MLST (Maiden, 2006)]. Although these methods are often well established, fast and comparatively cheap, their main drawback is lack of discriminatory power when it comes to typing of closely related isolates, for example isolates from a single outbreak of a bacterial pathogen. Many isolates, especially within a high-incidence setting, show an identical result with the fingerprinting methods, and have the same ‘type’ assigned. This prevents the definition of precise relationships between these isolates, and prohibits the identification of source cases or environmental sources, and an understanding of the detailed molecular architecture of bacterial epidemics.

The advent of comparatively cheap whole-genome sequencing technologies (next-generation sequencing) in the last few years seems to offer an easy solution, as these techniques monitor all changes in a bacterial genome, and therefore provide the maximum possible discriminatory power between two isolates. Such changes include single-nucleotide polymorphisms (SNPs) and small insertions or deletions (indels). Several recent studies have explored the possibilities that genomics offers to bacterial typing (an overview is given in Table 1) and here we highlight some of the advances in this field.

Hospital infections

Outbreaks of infections with health-care-associated pathogens, such as Clostridium difficile, Acinetobacter baumannii and methicillin-resistant Staphylococcus aureus (MRSA) are prone to insufficient resolution with currently used typing techniques. Especially the precise relationships within spreading MRSA remain unclear because the multilocus-sequence type ST239 accounts for at least 90% of health-care-associated MRSA in large parts of the world, including China (Xu et al., 2009), Thailand (Feil et al., 2008) and Turkey (Alp et al., 2009). Classical genotyping methods offer little discriminatory power to subtype ST239 isolates. Harris and colleagues (2010) therefore used a next-generation sequencing platform to analyse 63 isolates of subtype ST239, consisting of a global collection (43 isolates) and a local collection from a hospital in Thailand within a 7-month time frame (20 isolates). The phylogenetic tree (Fig. 1) established from core genes of these isolates was complemented with isolation date and geographical origin. The tree shows a high degree of consistency with the geographic source. Intercontinental transmission events were detected, such as the re-introduction of MRSA in Portuguese hospitals.
that must have originated from a South American variant, or a Danish isolate that clustered with the Thai clade. Patient records indicated that this Danish patient in question was actually a Thai national.

In addition to detecting intercontinental spread, this kind of fine-scale analysis holds the promise to detect transmission events within a single hospital. Five of the isolates from the Thai hospital were closely related to each other and suggested an epidemiological link between the respective patients. These patients were located in wards in adjacent blocks, in contrast to other patients with more divergent isolates. Such information is invaluable for interventions to target MRSA transmission.

In the UK, military patients returning from Iraq or Afghanistan are often colonized with multidrug-resistant A. baumannii (MDR-Aci) (Lewis et al., 2010). During an outbreak in 2008, four military patients were diagnosed with MDR-Aci infections, and subsequently two civilian patients were found to be colonized as well (Lewis et al., 2010). The application of next-generation sequencing shed light on transmission events within the outbreak, while standard typing techniques were unable to differentiate between alternative epidemiological hypotheses. Although a conservative SNP detection approach was chosen, the three identified SNPs were sufficient to detect transmission events within this small-scale outbreak.

Environmental sources and food-borne pathogens

If the source of a disease is a ubiquitous environmental source such as contaminated water, or bacterial spores that survive on nearly every surface, identification of the exact source might be impossible. Following the dynamics of an outbreak can become more important, such as for example for group A Streptococcus (GAS). Epidemics of GAS with an M3 serotype have an unusual periodicity of infection peaks of 4–7 years (Kohler et al., 1987; Colman et al., 1993). Although the currently used typing techniques allowed to establish a model of these recurring epidemics (Fig. 2), the full molecular complexity of the successive bacterial epidemics was only appreciated after performing a next-generation sequencing study (Beres et al., 2010). Sequencing of 95 isolates allowed the identification of a unique genome sequence for each isolate.

However, the still relatively high costs for next-generation sequencing makes it necessary to find other solutions if hundreds of strains need to be investigated. Many studies therefore apply (a subset) of their newly identified SNPs to additional isolates. The presence/absence patterns of these SNPs define a SNP type for each isolate. Clustering of the types allows the identification of groups with the same or a similar SNP type. This Table 1. Examples of genomic tracing of disease epidemics.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (Mb)</th>
<th>Disease Mode of transmission</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methicillin-resistant Staphylococcus aureus (MRSA)</td>
<td>1.9</td>
<td>Hospital infections Health-care associated hands</td>
<td>Harris et al. (2010) WGS</td>
</tr>
<tr>
<td>Multidrug-resistant Acinetobacter baumannii (MDR-Aci)</td>
<td>3.03</td>
<td>Hospital infection Contaminated clothing and bedding, bed rails, ventilators, sinks, door handles, carved or worn-appearing surfaces, detectors, medical devices</td>
<td>Lewis et al. (2010) WGS</td>
</tr>
<tr>
<td>Group A Streptococcus (GAS)</td>
<td>1.89</td>
<td>Hospital infection e.g., scarlet fever, pharyngitis</td>
<td>Beres et al. (2010) WGS</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>2.81</td>
<td>Listeriosis Contaminated clothing and bedding, food or water or inhalation of dust or cutaneous contact with beds or door handles</td>
<td>Gilmore et al. (2010) WGS and SNP typing</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>4.02</td>
<td>Tuberculosis Cutaneous contact with spores, inhalation of contaminated food or ingestion of contaminated food</td>
<td>Schürch et al. (2010a, b) WGS and SNP typing</td>
</tr>
<tr>
<td>Francisella tularensis</td>
<td>1.89</td>
<td>Tularaemia Scratches or bites from animals, consumption of contaminated meat or water or inhalation of spores</td>
<td>Pandya et al. (2008) Resequencing array and SNP typing</td>
</tr>
</tbody>
</table>

WGS, whole-genome sequencing.

© 2010 The Authors
Journal compilation © 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, Microbial Biotechnology, 3, 628–633
Fig. 1. Phylogenetic evidence for intercontinental spread and hospital transmission of health-care-associated MRSA isolates, type ST239. Maximum-likelihood phylogenetic tree based on core genome SNPs of ST239 isolates, annotated with the country and year of isolation. The continental origin of each isolate is indicated by the colour of the isolate name: blue, Asia; black, North America; green, South America; red, Europe; and yellow, Australasia. Bootstrap values are shown below each branch, with a star representing 100% bootstrap support. The scale bar represents substitutions per SNP site. A cladogram of the Thai clade is displayed for greater resolution with bootstrap values (above the branch), number of distinguishing SNPs (below the branch), and isolates labelled with date of isolation, where known. Reprinted from Harris et al. (2010), with permission from American Association for the Advancement of Science (AAAS).

Fig. 2. Model summarizing changes in group A Streptococcus (subclone M3) over time. The frequency distribution of all strains in the three epidemics is shown in grey, with three peaks of infection centred around 1995, 2000 and 2005. Ten major subclones (SC-1 to SC-10) were identified among the 344 strains collected from 1992 through 2007 based on different DNA-typing techniques. The widths of the coloured SC symbols show the temporal distribution of the SCs, and the heights are proportional to the annual abundance. Arrows between SCs indicate estimated relationships and give differences found in the loci assessed. The total number of isolates per year is given above the time line at the bottom. Reprinted from Beres and colleagues (2010). Copyright of the National Academy of Sciences.

© 2010 The Authors
Journal compilation © 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, Microbial Biotechnology, 3, 628–633
strategy has its own problems because it leads to branch collapse and linear phylogenies (Pearson et al., 2009; Beres et al., 2010). In the study of Beres and colleagues however, it allowed the identification of a complex population structure with micro- and macro-bursts of emerging clones (Beres et al., 2010).

For food-borne pathogens such as *Listeria monocytogenes*, quick identification of sources of infections is desirable. *Listeria monocytogenes* is ubiquitously present in our environment, and outbreaks are often caused by contaminated food such as milk, soft cheese, hot dogs and other processed foods. If *L. monocytogenes* is introduced into food-processing facilities, it can persist for a long time, as it is able to grow in refrigerated food (Ramaswamy et al., 2007). To track the sources of an outbreak, typing of the bacterial isolates of diseased patients and of potential sources is necessary. Two *L. monocytogenes* isolates of a large Canadian outbreak of listeriosis that was associated with ready-to-eat meat products were subjected to next-generation sequencing and the sequences compared (Gilmour et al., 2010). The identified SNPs, three indels and a prophage were then used to type other isolates of the same outbreak. The resulting evolutionary model is illustrated in Fig. 3, where isolates with an identical type cluster at the same nodes. This analysis indicated that three distinct strains were involved in the outbreak, and it was possible to study the strain-specific features of these outbreak strains.

**Human-to-human transmission**

Most infections of tuberculosis in humans result in asymptomatic, latent infections, and only about one in 10 infections progress to active disease. This can happen at any time in a patient’s life, which makes it often impossible to track the source of infection that might have been a contact of decennia ago. However, patient interviews can give some indications and this information was used when selecting three bacterial isolates for next-generation sequencing that were part of well-characterized transmission chains of a tuberculosis outbreak in the Netherlands (Schürch et al., 2010a,b). All other *Mycobacterium tuberculosis* isolates of the same outbreak were typed with the identified SNPs. By integration of SNP types, isolation dates and contact information, a detailed scheme of the outbreak was established (Fig. 4), and new transmission chains were identified. The study results comprised a surprising amount of information detail, such as the example of a married couple that both were infected with *M. tuberculosis* by a third source. Later, after the isolate underwent a single-nucleotide change, the couple infected each other. Furthermore, the genomic variability

---

**Fig. 3.** Evolutionary model for the *Listeria monocytogenes* isolates recovered during a nation-wide food-borne outbreak. Predicted mutational events are indicated on the diagonal lines, genotypes of the resulting lineages are denoted within circles, and isolates representative of those lineages are indicated to the right of solid dots. Sequenced isolates are denoted with bold text. Reprinted from Gilmour and colleagues (2010).
within populations of the same patient was addressed in this study, which can be considerable in *M. tuberculosis* isolates of the same patients (Al-Hajoj et al., 2010).

**Biological weapons**

Despite the widespread use of antibiotics, bacterial biological weapons remain a challenge to global security, especially with regard to bioterrorism. Tularaemia for example, caused by *Francisella tularensis*, is not a very common disease. However, its inclusion in biological warfare programmes (Dennis et al., 2001) makes the bacteria an interesting subject to study by next-generation sequencing (Pandya et al., 2009). Anthrax, an infamous biological warfare agent caused by *Bacillus anthracis*, was released by the Aum religious cult in Japan in 1993. The 2001 US-Anthrax attacks, where letters with infectious anthrax were delivered, caused the death of five people. It also underpinned the growing importance of identification of *B. anthracis* at the strain level for forensic investigations and source tracing (Chen et al., 2010; Segerman et al., 2010). Next-generation sequencing of two Japanese isolates (Kuroda et al., 2010) and the development of SNP assays enabled the discrimination of clusters and subgroups of isolates, and will aid in traceability of future anthrax bioterrorism attacks, at least if these are conducted with a known *B. anthracis* strain.

**Future developments**

In order to save lives through tracing of infectious diseases, it is necessary to discriminate isolates at the strain level. Next-generation whole-genome sequencing of bacterial isolates aids in identification of a source of an outbreak, determination of transmission events or description of the dynamics of an outbreak. Therefore, whole-genome sequencing should eventually replace or amend other bacterial typing methods in (clinical) microbiological laboratories.

However, although the future application of whole-genome sequencing is highly desirable, in order to achieve this in routine laboratory settings, the sequencing techniques and data analysis and storage need to be more efficient and come at lower costs, especially if used for thousands and thousands of strains. The quality and per-sample costs of the next wave of DNA sequencers that is expected in coming years will show us if this inevitable development will be accomplished in the near future.

**Acknowledgements**

We thank Kristin Kremer for critically reading and correcting the manuscript. R.S. is supported by the Netherlands Centre for Bioinformatics, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

**References**


