From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases

M J Struelens (Marc.Struelens@ecdc.europa.eu)[1], S Brisse[2]
1. European Centre for Disease Prevention and Control (ECDC), Stockholm, Sweden
2. Institut Pasteur, Paris, France

Citation style for this article:
Struelens MJ, Brisse S. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. Euro Surveill. 2013;18(4):pii=20386. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20386

Article published on 24 January 2013

The use of increasingly powerful genotyping tools for the characterisation of pathogens has become a standard component of infectious disease surveillance and outbreak investigations. This thematic issue of *Eurosurveillance*, published in two parts, provides a series of review and original research articles that gauge progress in molecular epidemiology strategies and tools, and illustrate their applications in public health. Molecular epidemiology of infectious diseases combines traditional epidemiological methods with analysis of genome polymorphisms of pathogens over time, place and person across human populations and relevant reservoirs, to study host–pathogen interactions and infer hypotheses about host-to-host or source-to-host transmission [1-3]. Based on discriminant genotyping of human pathogens, clonally derived strains can be identified as likely links in a chain of transmission [1-3]. In this two-part issue of *Eurosurveillance*, Goering et al. explain that such biological evidence of clonal linkage complements but does not replace epidemiological evidence of person-to-person contact or common exposure to a potential source [3]. Muellner et al. provide clear examples how prediction about infectious disease outcome and transmission risks can be enhanced through integration of pathogen genetic information and epidemiological modelling to inform public health decisions about food-borne disease prevention [4].

As reviewed by Sabat et al., epidemic source tracing requires timely deployment of high resolution typing methods that index variation of genomic elements with a fast molecular clock [1-5]. For outbreak studies, comparative methods, as opposed to library typing methods, are sufficient, and the higher the power to resolve micro-evolutionary distance, the greater the likelihood to decide between alternative transmission hypotheses generated by observational epidemiology [1-6]. Once standardised to enable a uniform genotype nomenclature across laboratories, thereby providing a library typing system, such discriminatory methods can be further applied to control-oriented surveillance [1-5]. Early outbreak detection is achieved by genotyping prospectively as many consecutive cases in a population as possible to identify clusters of clonally linked isolates [5]. Examples include PulseNet, the nationwide food-borne disease surveillance system in the United States [7] as well as national molecular surveillance schemes developed to detect clusters of tuberculosis as described by Fitzgibbon et al. [8]. Library typing systems that use more stable genotypic markers such as bacterial multilocus sequence typing (MLST) are suitable for strategy-oriented molecular surveillance aimed at monitoring secular trends in the evolution of pathogen genotypes and in their distribution over larger geographic and population scales [1-5]. Such molecular surveillance systems can call attention to the emergence of strains with enhanced virulence or drug resistance, help identify risk factors associated with transmission of specific strains, or predict the effectiveness of public health measures such as vaccinations. This approach is well established for global virological surveillance of human and avian influenza. As illustrated by an experience from New-Zealand presented by Muellner et al., a nationwide molecular surveillance of campylobacteriosis using a sequential combination of typing systems can inform both disease control measures and prevention policies by detecting local outbreaks and modelling endemic disease attribution to specific food sources [4]. Structured surveys that combine spatiotemporal mapping of strain genotype and antimicrobial resistance phenotype is a powerful means to monitor the emergence and spread of multidrug-resistant clones across a continent, as reported by Chisolm et al. for *Neisseria gonorrhoeae* in Europe [9].

As summarised by Sabat et al., there have been continuous technological improvements for microbial genomic characterisation in the past decade, moving from fingerprinting methods such as pulsed-field gel electrophoresis of bacterial macrorestriction fragments to more robust, portable and biologically informative assays such as bacterial multilocus variable-number tandem repeat analysis (MLVA) and sequencing of single/multiple loci of both bacterial and viral human pathogens [3-5,9-11]. With the decreasing cost and continuing refinement of high-throughput

genome sequencing technologies, we are now witnessing a quantum leap from genotypic epidemiology to genomic epidemiology as whole viral or bacterial genomes become open to scrutiny at population level. As reviewed by Carrico et al., advances in laboratory typing tools have been enabled by parallel progress in the information technology needed to capture genetic data on pathogens, and in quality control, formatting, storage, management and, most importantly, bioinformatics analysis and real-time electronic data sharing through online databases [10].

Among the sequence-based genotyping assays, MLST is widely applied for epidemiological investigations of bacterial and fungal pathogens and is a primary typing method for clonal delineation in pathogens such as *Neisseria* [12] or *Campylobacter* [4]. The advantages of MLST are twofold: firstly, it generates reproducible and standardised data that are highly portable (i.e. easily transferrable between different systems) and comparable across laboratories in centralised databases accessible through the Internet. Secondly, the nucleotide substitutions that underlie MLST variation can be interpreted directly in terms of population genetics and evolutionary processes. Because nucleotide polymorphisms evolve slowly in bacteria, MLST is very appropriate to describe the patterns of genetic variation within bacterial species at the global scale. Therefore, one of the major applications of MLST is to decipher bacterial population structure, including clonal diversity, to create a phylogenetic structure of different lineages and to assess the impact of homologous recombination. Recently, this has led to a bold proposal to replace the 70 year-old serotyping nomenclature system for *Salmonella* strains with MLST [13].

To reduce costs and increase speed, typing based on the sequencing of single highly variable genes was developed for a few pathogens. The most widely used systems are sequencing of the *emm* gene coding for the M antigen of *Streptococcus pyogenes* (which can be compared to the results from traditional M serotyping) and the *spa* gene coding for surface protein A of *Staphylococcus aureus* [5]. However, single locus typing approaches are limited by events such as homoplasy (evolutionary reversion or convergence) and horizontal gene transfer, as discussed by Sabat et al. [5].

Lindstedt et al. show in this issue how interest in MLVA has grown from the limitations of MLST and other methods to discriminate among isolates of epidemiologically important clones, such as *Escherichia coli* O157:H7 and *Salmonella* serovar Typhimurium [11]. MLVA retains the 'multilocus' concept of MLST but is based on rapidly evolving loci characterised by the presence of short, tandem repeated sequences. MLVA has proven very useful in surveillance and epidemiology, e.g. for monitoring clonal trends, cluster detection and outbreak investigation [5,11,14]. The high discriminatory power of MLVA for many bacterial groups, combined with its simplicity, makes it an especially useful subtyping tool for so-called monomorphic pathogens [5,11]. In addition, MLVA has a strong potential for inter-laboratory standardisation, and several web-accessible database systems have been developed [5,10-11]. One important drawback is that many MLVA schemes are highly specific for given clones, thus limiting their applicability. Furthermore, for long-term epidemiology or population biology, MLVA markers can be affected by homoplasy, which renders MLVA data less robust than MLST as a library typing system and for phylogenetic purposes. It also remains unclear whether assembly of high throughput sequence data will be reliable enough to determine MLVA alleles, as the repeat arrays pose particular technical challenges for current high throughput sequencing technologies.

From a perspective of medical and public health microbiology and epidemiology, whole genome sequencing (WGS) combines two decisive advantages compared to previous methods: it provides maximal strain discrimination on the one hand, and can be linked to clinically and epidemiologically relevant phenotypes on the other hand. The method is widely seen as the ultimate tool for epidemiological typing of bacteria and other pathogens. It has already proven highly informative to resolve local S. aureus outbreaks [6] as well as elucidate the evolutionary events leading to the emergence and global dissemination of super-pathogen clones with enhanced virulence and multidrug resistance, such as Clostridium difficile ribotype 027 strains [14-15]. Moreover, WGS will provide full genomic characteristics of the infectious isolates, including the set of genes linked to antimicrobial resistance (the resistome) and those linked to virulence of the isolates (the virulome). As discussed by several authors in this issue [3,5,10,12,14], WGS still remains to be fully harnessed conceptually and fine-tuned technologically. This promising technology currently faces three major challenges: speed, data analysis and interpretation, and cost.

As opposed to previous sequence-based typing methods, WGS will change the way we look at pathogen diversity in one fundamental way: without an a priori focus on a subset of loci. As all genetic information will be available, it will allow the discovery of novel, unexpected variation, including polymorphisms that evolve during outbreaks or changes that are selected in vivo during infection. Such pathoadaptive changes can result in increased virulence or novel pathophysiological processes. One example of such a micro-evolutionary change is the emergence during influenza A(H1N1)pdm09 epidemic of a quasispecies variant with a haemagglutinin D222G mutation which is associated with modified tissue receptor tropism and severe influenza virus infections, as reported by Rykkvin et al. in this journal [16]. Due to the rapid rate of evolution of viruses and their small genomes, virologists have long been using genome-wide sequencing. The term 'phylodynamics' designates the study of the interplay of epidemiological and evolutionary patterns, pioneered in

virology [17]. Phylodynamics based on WGS of bacterial populations is emerging as a fertile field of investigation for public health microbiology [5-6,14-15].

As discussed by Jolley and Maiden, WGS sequencing of bacterial pathogens and archiving of the collected data will raise the issue of genomic strain nomenclature [12]. One particularly interesting advantage of MLST in the era of high-throughput sequencing lies in its forward compatibility with future whole genome sequencing, or core genome allotyping, as underlined by Sabat et al. and Jolley and Maiden [5,12]. Several recent tools allow extracting MLST information from high-throughput sequencing data [12,18,19]. The BIGSDB bioinformatics application incorporates MLST databases and provides the possibility to extend the MLST approach to include the full core genome [12]. We anticipate that a WGS-based genotype nomenclature could be developed as a complement to the well-established MLST nomenclature of bacterial clones. As core genome evolution within MLST clones is mainly mutational, the possibility to reconstruct phylogeny based on WGS data should allow a hierarchical classification of WGS types, giving access to different levels of genetic distance resolution depending on the epidemiological questions and length of the study period. This is just one example of the challenges that we face as we enter the exciting era of genomic epidemiology [5,10,12].

Beyond the hurdles in technology and bioinformatics that we still need to overcome, what are the needs for translating advances in genomic epidemiology into public health benefits? Laboratory-based surveillance is pivotal to monitoring infectious disease threats to human health. It relies on aggregating microbiological data that are produced at clinical care level and supplemented by reference laboratory testing. As highlighted by Niesters et al., molecular methods supplant culture-based diagnostic methods, thereby making genomic information relevant to disease surveillance available at the level of the diagnostic laboratory. This technological shift challenges the hierarchical architecture of surveillance networks that relies on samples and culture specimens being referred from the clinics to the reference laboratories and public health institutes [20]. Niesters et al. describe the pilot experience with the TYPENED surveillance network as a molecular data-sharing platform pioneered in the Netherlands by a consortium of clinics, academic institutions and public health virology laboratories [20]. This collaborative approach led to a consensus on how to choose surveillance targets, harmonise sequence-based virological diagnostic assays and share sequence data through a common platform [20].

In addition to stimulating changes in public health systems, the application of high-resolution typing tools such as WGS in outbreak management raises a number of ethical questions, as discussed by Rump et al. in this journal [21]: protection of personal data, informed consent with regard to the investigation of clinical samples, and moral responsibility and legal liability to act upon the evidence to prevent or mitigate disease transmission. As real-time data sharing becomes technically feasible for surveillance and cross-border outbreak investigations, public health organisations will need to develop a policy for the use of these data that balances risks and benefits and defines adequate governance. As part of its mandate to foster collaboration between expert and reference laboratories supporting prevention and control of infectious diseases, the European Centre for Disease Prevention and Control (ECDC) is facilitating interdisciplinary collaboration and assessing public health needs for the integration of microbial genotyping data into surveillance and epidemic preparedness at European level [22]. As announced recently, a European data exchange platform that combines typing data with epidemiological data on a list of priority diseases is being piloted for molecular surveillance of multidrug-resistant *Mycobacterium tuberculosis* and food-borne pathogens [23]. As WGS gradually becomes part of epidemiological studies, ECDC is party to the international expert consultations aimed at building interoperable databases of microbial genomes for future application in public health [24].

References

1. Struelens MJ, De Gheldre Y, Deplano A. Comparative and library epidemiological typing systems: outbreak investigations versus surveillance systems. Infect Control Hosp Epidemiol. 1998;19(8):565-9.
2. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13 Suppl 3:1-46.
3. Goering RV, Köck R, Grundmann H, Werner G, Friedrich AW, on behalf of the ESCMID Study Group for Epidemiological Markers (ESGEM). From Theory to Practice: Molecular Strain Typing for the Clinical and Public Health Setting. Euro Surveill. 2013;18(4):pii=20383. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20383
4. Muellner P, Pleydell E, Pirie R, Baker MG, Campbell D, Carter PE, et al. Molecular-based surveillance of campylobacteriosis in New Zealand – from source attribution to genomic epidemiology. Euro Surveill. 2013;18(3):pii=20365. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20365
5. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013;18(4):pii=20380. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20380
6. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012;366(24):2267-75.
7. Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, et al. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. Foodborne Pathog Dis. 2006;3(1):36-50.
8. Fitzgibbon MM, Gibbons N, Roycroft E, Jackson S, O'Donnell J, O'Flanagan D, et al. A snapshot of genetic lineages of Mycobacterium tuberculosis in Ireland over a two-year period, 2010 and 2011. Euro Surveill. 2013;18(3):pii=20367. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20367
9. Chisholm SA, Unemo M, Quaye N, Johansson E, Cole MJ, Ison CA, et al. Molecular epidemiological typing within the European Gonococcal Antimicrobial Resistance Surveillance Programme reveals predominance of a multidrug-resistant clone. Euro Surveill. 2013;18(3):pii=20358. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20358
10. Carriço JA, Sabat AJ, Friedrich AW, Ramirez M, on behalf of the ESCMID Study Group for Epidemiological Markers (ESGEM). Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill. 2013;18(4):pii=20382. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20382
11. Lindstedt BA, Torpdahl M, Vergnaud G, Le Hello S, Weill FX, Tietze E, Malorny B, Prendergast DM, Ní Ghallchóir E, Lista RF, Schouls LM, Söderlund R, Börjesson S, Åkerström S. Use of multilocus variable-number tandem repeat analysis (MLVA) in eight European countries, 2012. Euro Surveill. 2013;18(4):pii=20385. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20385
12. Jolley KA, Maiden MC. Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill. 2013;18(4):pii=20379. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20379
13. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, et al. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. PLoS Pathog. 2012;8(6):e1002776.
14. Knetsch CW, Lawley TD, Hensgens MP, Corver J, Wilcox MW, Kuijper EJ. Current application and future perspectives of molecular typing methods to study Clostridium difficile infections. Euro Surveill. 2013;18(4):pii=20381. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20381
15. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet. 2012;45(1):109-13.
16. Rykkvin R, Kilander A, Dudman SG, Hungnes O. Within-patient emergence of the influenza A(H1N1)pdm09 HA1 222G variant and clear association with severe disease, Norway. Euro Surveill. 2013;18(3):pii=20369. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20369
17. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303(5656):327-32.
18. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012;50(4):1355-61.
19. Inouye M, Conway TC, Zobel J, Holt KE. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics. 2012;13:338.
20. Niesters HG, Rossen JW, van der Avoort H, Baas D, Benschop K, Claas EC, Kroneman A, van Maarseveen N, Pas S, van Pelt W, Rahamat-Langendoen JC, Schuurman R, Vennema H, Verhoef L, Wolthers K, Koopmans M. Laboratory-based surveillance in the molecular era: the TYPENED model, a joint data-sharing platform for clinical and public health laboratories . Euro Surveill. 2013;18(4):pii=20387. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20387
21. Rump B, Cornelis C, Woonink F, Verweij M. The need for ethical reflection on the use of molecular microbial characterisation in outbreak management. Euro Surveill. 2013;18(4):pii=20384. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20384
22. Palm D, Johansson K, Ozin A, Friedrich AW, Grundmann H, Larsson JT, et al. Molecular epidemiology of human pathogens: how to translate breakthroughs into public health practice, Stockholm, November 2011. Euro Surveill. 2012;17(2):pii=20054. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20054
23. van Walle I. ECDC starts pilot phase for collection of molecular typing data. Euro Surveill. 2013;18(3):pii=20357. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20357
24. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, Harmsen D, et al. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis. 2012;18(11):e1