Public Health Genomics

Public Health Genomics 2013;16:25–30
DOI: 10.1159/000342709

Published online: March 18, 2013

Public Health Genomics and the New Molecular Epidemiology of Bacterial Pathogens

M.W. Gilmour M. Graham A. Reimer G. Van Domselaar

National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Man., Canada

Key Words
Genome sequencing · Molecular typing · Outbreak investigation · Public health

Abstract
Laboratory methods that can unambiguously fingerprint pathogenic microbes are needed to investigate the transmission of human infectious diseases from diverse sources, such as from the community, from the environment, within hospitals, or from contaminated food or water sources. Public health investigations currently rely on laboratory subtyping methods that ultimately provide only a fraction of the total genetic information of a pathogen, and although there is widespread success using existing subtyping methods, they do not always provide sufficient evidence to link disease cases together into outbreaks or to link these human cases to the culprit source. Alternatively, whole-genome sequencing of bacterial pathogens provides an unabridged examination of the genetic content of individual pathogen isolates, enabling public health laboratories to benefit from comparative analyses of total genetic content. In this context, whole-genome sequencing represents the ultimate epidemiological typing method – a universally applicable, highly detailed typing platform capable of providing the entire genetic blueprint of a pathogen and distinguishing strains to the single nucleotide level. These new genomic methods, if implemented within existing public health laboratory response programs, promise to revolutionize the ability of the laboratory to provide information and evidence on the evolution, transmission and virulence for bacterial pathogens – and this revolution is launching the new field of '*genomic epidemiology*'.

The 'Genomic Epidemiology' Era Has Arrived

The manner in which laboratories acquire evidence on the microbial identity and fingerprint subtypes of human pathogens needed for the treatment, monitoring and response to infectious disease is on the cusp of a major transformation. In the past, whole-genome sequencing was exclusively applied as a research tool, but recent technological advances have made it feasible and even advantageous for use during public health investigations of bacterial disease events. These innovations in DNA sequencing technologies are now providing investigators with the complete catalogue of a pathogen's genome (instead of just a small fraction of it), delivered within a cost and time frame that is approaching relevancy for clinical and public health laboratories. Inherent in a genome is a patho-

Matthew W. Gilmour
National Microbiology Laboratory
1015 Arlington Street
Winnipeg, MB R3E 3R2 (Canada)
E-Mail Matthew.Gilmour@phac-aspc.gc.ca

Matthew W. Gilmour
Diagnostic Services of Manitoba, Clinical Microbiology
MS-673C 830, Sherbrook Street
Winnipeg, MB R3A 1R9 (Canada)
E-Mail mgilmour@dsmanitoba.ca

KARGER

E-Mail karger@karger.com
www.karger.com/phg

1662–4246/13/0162–0025$38.00/0

gen's identity, but also available for discovery are all other informative markers that could represent the classical typing designations such as serotype as well as pathogenic determinants including toxin type, antimicrobial resistance genes and host-adhering proteins. Historically, such pieces of evidence have been acquired through multiple, independent laboratory tests; now they are present en masse within a genome sequence. Recent examples demonstrating the speed and robustness of genomic-based approaches to study the emergence of a priority bacterial pathogen were seen during the response to the 2011 European outbreak of toxigenic *Escherichia coli* O104:H4 [1–3].

Whole-genome sequences provide the most fulsome data set for pathogen relatedness and diversity studies, enabling thorough 'molecular epidemiology' investigations to decipher the transmission network of pathogens. In such a process, clusters of human illness are detected via joint analysis of pathogen subtyping and case epidemiologic data in an attempt to discover and track common microbial fingerprint 'matches' between human isolates and those isolated from potential sources. Laboratory technologies such as pulsed-field gel electrophoresis (PFGE) and multi-locus sequence typing (MLST) are some of the current laboratory 'gold standards' applied to characterize and subtype human-clinical, animal, food, and environmental isolates; these methods sample only a limited proportion of the total genetic information. With these lower resolution methods, it is not always possible to accurately detect clusters of human disease, nor to properly attribute source and conduct trace-back analyses between human disease occurrences by matching subtypes to the candidate sources that caused human disease. For example, some pathogens such as *Salmonella* serovar Enteritidis (SE) appear highly homogeneous (i.e. a small number of clones predominate), and therefore, the current typing techniques lack sufficient discriminatory power to distinguish event-related from unrelated or endemic isolates. As such, 'matching' between SE subtypes is overly common to the point where little significance can be assigned to any given match, and the incidence of the highly predominant subtypes grows alongside the continued surveillance and outbreak detection efforts. Consequently, with a lack of sufficient information on the true genetic diversity within a given pathogen population, many probable outbreaks go undetected and sources of SE infection are frequently not identified. The benefits of whole-genome based examinations into an endemic pathogen clone to reveal the true levels of genetic diversity (and transmission patterns) have, for example, been demonstrated for *emm*59 type strains of group A *Streptococcus* [4, 5].

Genome sequencing overcomes limitations on the scope of information that can be collected for each pathogen isolate, and pathogen-wide analyses offer the potential to uncover informative high-resolution markers (such as the precise location and distribution of single nucleotide polymorphisms). Less than a decade ago, the sequencing of a single pathogen strain required months to years and hundreds of thousands of dollars with a large team to complete. Today, with advanced DNA sequencing platforms that can produce gigabases of data, this same work can be completed in a matter of days, using a much smaller team and a few hundred dollars per strain [6]. Owing to these technological improvements and cost savings, a relevant scenario under which to apply the burgeoning genome sequencing technologies is to support (or lead) public health investigations. None of the currently used subtyping methods can equal the theoretical discriminatory power of DNA sequencing technologies that provide single base resolution spanning the entire genome. With such technology, investigators are not reliant on subtyping platforms that examine a single gene or a small number of markers (be they 'house keeping' loci in MLST, tandem repeats or other features along the chromosome); genome sequencing can comprehensively gather substantial biological data within a single experiment. Accordingly, genome-based molecular epidemiology, or rather *genomic epidemiology*, is rapidly becoming a bona fide analytic approach applied during public health investigations.

Genome Databases Require Expansion for Public Health Purposes

Contemporary typing methods such as PFGE and MLST have allowed the formation of international outbreak detection networks and/or have standardized the approaches for pathogen phylogeny and population studies. If genomic epidemiology is to achieve similar widespread adoption as a public health application, it will be necessary to draw from many of the guiding principles that led to the acceptance and broad use of the contemporary typing methods. These include: method standardization and interoperability between labs; centralization and accessibility of subtype databases with uniform assessment parameters; and importantly, a depth and breadth of testing coverage across a pathogen population to allow for proper epidemiological investigation.

A current strength of applying PFGE or MLST for outbreak detection is the amount of typing information available in centralized databases that can be used to compare the subtypes acquired for contemporary isolates relative to historical isolates (be they from human isolates or other sources). Public health networks such as PulseNet house thousands of PFGE fingerprint entries for pathogens such as *Escherichia coli* O157 and *Listeria monocytogenes (Lm)*; as new cases or surveillance isolates are detected, meaningful comparison of their laboratory fingerprints can occur against the voluminous database of historical fingerprints collected from across multiple countries and multiple sources (human, animal, food and environmental). Such robust databases increase the likelihood of identifying significant matches, or alternatively, permit the exclusion or diminished prioritization of insignificant matches, as in the case of highly predominant subtypes.

This same depth of strain-to-strain comparisons cannot yet occur for pathogen whole-genome sequences. The reason: despite collecting detailed genetic information per isolate via sequencing, the current paucity of pathogen genome sequences means that essential information for genomic epidemiologic assessments and interpretation are missing. For example, our team sequenced 2 clinical isolates of *Lm* during a nationwide foodborne outbreak in Canada in 2008 [7]. Although this *Listeria* genomics work was a success from the standpoint of a technological milestone (being one of the first to sequence genomes during a public health event) [8], our ability to interpret this genomic data set within the context of genomic epidemiology was severely limited by a scarcity of other available *Lm* genome sequences for comparative genomic examination. In 2008, we were restricted to broad conjectures on the evolution and transmission of *Lm* strains during the listeriosis outbreak based on the traits identified in our 2 representative strains by necessity. Our analyses would have been more informative if it had been feasible to obtain additional genome sequences from isolates sampled broadly across the food production environment, retail meat products and the human-clinical continuum. Moreover, if a larger stock of existing genomic data had existed for recent and historical Canadian *Lm* strains, analysis and modeling of the particular outbreak strains could have had a deeper and more accurate context.

These same issues were encountered during multiple genomic studies to investigate the origins of the cholera epidemic that began in Haiti in 2010 following its devastating earthquake [9]. Multiple groups compared Haitian *Vibrio cholerae (Vc)* genomes to other *Vc* genomes; these studies either followed up on epidemiologic inferences (comparing Haitian strains to concurrent *Vc* isolates from Nepal) [10] or laboratory inferences (comparing Haitian strains to global and historical *Vc* isolates that had matching PFGE patterns) [11, 12]. These authors all similarly concluded that in order to achieve accurate source attribution (especially for an organism known to be involved in pandemics and intercontinental transmission), it is necessary to sample as broadly as possible to obtain a true representation of the global *Vc* population. Simply put, if only a small proportion of the global *Vc* population is sampled then even a near-identical genetic match between 2 locales may not be highly significant, as additional and closer matches may exist in heretofore unsampled locales.

An unbiased sampling of the global strain diversity and genetic content is thus required to serve as the 'denominator' for any molecular epidemiologic investigation. For example, it is important to avoid sampling too narrowly; otherwise, one may err in extrapolating from a sample that represents a minor fraction of the true population. Through broad sampling (as done for cholera by Mutjera et al. [12]), researchers gain accurate understanding of the true scope of genetic diversity and obtain an accurate map of the distribution of specific traits/markers within a pathogen population, each of which in turn can be applied to track evolution and transmission of pathogens. Sampling strategies to generate informative genome databases need to include environmental, sporadic and outbreak-associated isolates that are temporally distributed (historical, contemporary and prospective) and also geographically distributed (local, national, global, and travel-associated). Sampling should also be informed by applying other phylogenetic information, such as MLST, to ensure the broadest possible representation is achieved across the known population. The associated metadata (source and dates of isolation, associated illness descriptors, etc.) for each strain are also crucial, as are any known clinical indicators of pathogenicity and virulence.

Whether the goal of a genomic investigation is direct trace-back of an individual human illness to a specific source, e.g. a specific lot of food that is contaminated; source attribution to more generic risks, e.g. food commodities that persistently contribute to disease; or the identification of markers of pathogenesis or niche persistence, the analyses will always be influenced by the number of input pathogen genomes. Continued expansion of whole-genome sequence data sets will therefore be necessary to accurately assess and trace infectious disease

events. With the newfound discriminatory power that genome sequencing offers, researchers still need to contextualize the diversity observed within outbreak events relative to the expected diversity for each pathogen. Only through the accumulated experience and data from sequencing and analyzing genomes across a broad spectrum of a pathogen population will sufficient baseline understanding be acquired to reconcile observed diversity (or perhaps lack thereof) in future genomic epidemiology studies of outbreaks.

Standardization of Pathogen Genome-Wide Analyses

Existing outbreak response systems, such as PulseNet, apply highly standardized typing methodologies and analysis metrics (and in most instances) under strict quality control and quality assurance. Interoperability between PulseNet laboratories is achieved via standardized protocols for acquisition of PFGE fingerprints, and robust comparisons of PFGE subtypes for the purposes of molecular epidemiology are completed through centralization of the functions related to fingerprint assignment and analysis of PFGE subtypes. As national activities for fingerprint pattern designation are completed, they are entered into centralized databases that house data contributed by all of the network partners. Similarly, once an assembled or partially assembled microbial genome data set is acquired, the resulting genome assembly represents a relatively standardized piece of public health intelligence that could be shared amongst collaborating parties [2]. However, for genomic epidemiology analysis to occur, data for other pathogen genomes must be available (discussed above) but with unified approaches for sequence assembly, feature annotations, associated metadata descriptions, and marker discover, which are currently lacking. Fortunately, important groundwork on conventions has been laid by other sequence-based consortiums [13] and discussions among the global microbiological community have initiated [14].

Decoding a microbial genome is a complex scientific task comprised of DNA sequence acquisition, sequence validation, biological interpretation, and functional analysis of features. Assembly of fragmented DNA sequence reads and analysis of their biological significance is comparable to solving a complicated puzzle made up of several million pieces. Fortunately, numerous successful technological advances and computational approaches have evolved. Once assembled, genome data sets provide a full characterization for emergent clones (including information regarding toxins, virulence determinants, antimicrobial resistance, or other novel traits). Importantly, they also provide comprehensive evidence to support the evolutionary relationships among strains. For the purposes of genomic epidemiology, genome-wide analyses can reveal informative markers (such as single nucleotide polymorphisms, genomic islands, or other insertions and deletions) that represent lineage-specific and virulence-related genetic traits that inform the microbe's true evolution and transmission history [15–18]. As such, with genomics, we are moving away from the simplification of subtyping 'integers' (wherein subtyping data such as MLST or PFGE patterns are represented by proxy as numerical designations) to a more detailed measure of strain relatedness using the full genetic content. Integer subtypes succeed in communicating strain subtype designations between investigators, however, if phylogenetic representations that contextualize the population of subtypes observed for that pathogen (e.g. dendograms, minimum-spanning trees) are not provided alongside the simplified integer designations, these classifications can actually mask the underlying genetic relatedness between subtypes and skew a full and proper interpretation. Although genomics can provide accurate representation of genetic content and diversity between strains, how this strain diversity, relatedness and evolution is uncovered and communicated is not yet standardized.

Bioinformatics approaches for pathogen-wide analyses, unlike the highly standardized analysis metrics of PulseNet, are highly varied across the global community, with an abundance of tools continually being developed, refined and packaged together as software pipelines. In order to support future, expanded and real-time studies investigating priority microbes, there is pressing need to streamline and formalize these data analysis workflows, such that pathogen sequence acquisition, assembly, feature annotation, and comparative pathogen analyses can readily identify epidemiologically significant traits within data sets in a timely manner. This is also not withstanding the time needed for biological interpretation specific to the problem and pathogen(s) at hand. Development of a functionalized bioinformatics pipeline (or agreement on analysis principles) needs to come from the core tenets of so-called 'open source' software: emphasizing *experimentation* and *accessibility*. As academic and public health researchers deploy a multitude of tools and approaches, those that have some measure of success to robustly analyze pathogen genome information will see more widespread use and will undergo subsequent refinement and

adoption. From there, implementation in more front-line public health sectors such as the regional reference and clinical laboratories could become a possibility.

Implementation of Genomics into Public Health Laboratories Is on the Horizon

Public health laboratories across the globe wish to acquire capacity, as well as experience, that will enable them to correctly interpret these comprehensive genomic data sets and to better support public health investigations. However, prior to implementation as a mainstream public health tool (and only later as a clinical method for patient management), whole-genome based approaches will have to evolve to include a centralized wealth of reference genomes for access by public health investigators, standardization of minimal data sets and analysis approaches, agreement on data interpretation and communication principles, and access to computational and capacity to process or interpret the data. These still remain outstanding issues at centres equipped with sequencing and bioinformatics infrastructures. Consequently, genomic epidemiology currently represents a research approach that can supplement, but not yet replace current gold standard typing methods and has to be refined and stabilized before widespread adoption in the public health field. However, it remains important to move in this direction.

For most public health and hospital laboratories, the prospect of performing whole-genome sequence analysis is not immediately achievable and is instead a medium-term goal. Large genome sequencing centres, national public health laboratories and academic groups continue to develop and make available genome analysis pipelines. With a growing number of applied genomic epidemiology examples, other public health laboratories have an opportunity to collaborate with these groups and begin applying these technologies to their pressing issues. The greatest opportunity to make the burgeoning genomic analyses platforms available to public health investigators may be through cloud-based systems [19, 20], wherein users do not need to localize computationally intensive analysis steps, but instead push their raw data to analysis systems curated by expert groups. Of note, even in the presence of cloud-based analysis applications, there will remain a requirement at each institute for continued skills and competency development in the application of genomics tools for purposes of public health investigations.

There exist other practical considerations beyond the availability of the requisite analysis pipelines that may pose barriers to the rapid adoption of genomic epidemiology in some public health labs. To be most beneficial, genome sequencing infrastructure must be readily available – requiring capital and support. For some, it may remain most economically feasible to perform whole-genome sequencing on only a small subpopulation and then to port genome-identified traits to higher throughput detection and typing technologies [21] in order to conduct routine screening for the distribution of these traits in wider isolate panels (i.e. those not yet sequenced). However, in light of emergent single-molecule DNA sequencing platforms [22] and the sheer scale of information returned with genomics, some public health labs may soon favour genomics over other conventional laboratory typing approaches (e.g. molecular subtyping or serotyping), for reasons of cost and availability of reagents, and perhaps the lack of small animals needed for antibody production. As such, some molecular typing platforms might soon be eclipsed by whole-genome sequencing as a practical reality, as it may already be more cost-effective to pursue a genomics-based approach rather than investing in largely proprietary higher throughput assay platforms.

Conclusion

Genomics is being positioned for widespread implementation as a public health tool for bacterial disease investigations due to the scale of pathogen genome information now available and the ability to share and transmit genome evidence between investigators. Before this can occur routinely, it is first necessary to reconcile the time and costs needed to complete such experiments as compared to the current laboratory identification and typing methods. Moreover, as investigators move from integer-based subtyping to the 'new analytical math' of full genomic epidemiology, there are new requirements for conventions and standard processes to store and mine genomic data as well as the need for sufficiently populated pathogen genome databases to complete robust and uniform genomic epidemiologic investigations. Fortunately, a bacterial genome represents an unbiased, unabridged capture of pathogen information, with the concerted efforts currently underway to develop both the technology and informatics requirements, it is easy to forecast into the near future that genomics will someday be used routinely in public health surveillance and outbreak response systems.

References

1 Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK: Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 2011;365:709–717.

2 Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, Wang J, Xu J, Pallen MJ, Wang J, Aepfelbacher M, Yang R; E. coli O104:H4 Genome Analysis Crowd-Sourcing Consortium: Open-source genomic analysis of Shiga toxin-producing *E. coli* O104:H4. N Engl J Med 2011;365:718–724.

3 Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H: Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 2011;6:e22751.

4 Fittipaldi N, Beres SB, Olsen RJ, Kapur V, Shea PR, Watkins ME, Cantu CC, Laucirica DR, Jenkins L, Flores AR, Lovgren M, Ardanuy C, Liñares J, Low DE, Tyrrell GJ, Musser JM: Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A *Streptococcus*. Am J Pathol 2012;180:1522–1534.

5 Fittipaldi N, Olsen RJ, Beres SB, Van Beneden C, Musser JM: Genomic analysis of emm59 group A *Streptococcus* invasive strains, United States. Emerg Infect Dis 2012;18:650–652.

6 Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW: A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open 2012;2:pii:e001124.

7 Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C: High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC Genomics 2010;11:120.

8 Baldry S: Attack of the clones. Nat Rev Microbiol 2010;8:390.

9 Centers for Disease Control and Prevention (CDC): Update on cholera – Haiti, Dominican Republic, and Florida, 2010. MMWR Morb Mortal Wkly Rep 2010;59:1637–1641.

10 Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM: Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio 2011;2:e00157–11.

11 Reimer AR, Van Domselaar G, Stroika S, Walker M, Kent H, Tarr C, Talkington D, Rowe L, Olsen-Rasmussen M, Frace M, Sammons S, Dahourou GA, Boncy J, Smith AM, Mabon P, Petkau A, Graham M, Gilmour MW, Gerner-Smidt P; V. cholerae Outbreak Genomics Task Force: Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. Emerg Infect Dis 2011;17:2113–2121.

12 Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JL, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G: Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 2011;477:462–465.

13 Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al: Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 2011;29:415–420.

14 Danish Technical University: Perspectives of a global, real-time microbiological genomic identification system – implications for national and global detection and control of infectious diseases. Consensus report of an expert meeting, Brussels, September 2011. http://www.food.dtu.dk/upload/fødevareinstituttet/food.dtu.dk/publikationer/2011/consensus%20report%20perspectives%20of%20a%20global,%20real-time.pdf.

15 Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR Jr, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R: Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 2011;43:1275–1280.

16 Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, Dougan G, Gordon MA: High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal *Salmonella* Typhimurium disease. Clin Infect Dis 2012;54:955–963.

17 Harris SR, Clarke IN, Seth-Smith HM, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, Lewis DA, Spratt BG, Unemo M, Persson K, Bjartling C, Brunham R, de Vries HJ, Morré SA, Speksnijder A, Bébéar CM, Clerc M, de Barbeyrac B, Parkhill J, Thomson NR: Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 2012;44:413–419.

18 Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: Evolution of MRSA during hospital transmission and intercontinental spread. Science 2010;327:469–474.

19 Kahn SD: On the future of genomic data. Science 2011;331:728–729.

20 Angiuoli SV, White JR, Matalka M, White O, Fricke WF: Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. PLoS One 2011;6:e26624.

21 Taboada EN, Ross SL, Mutschall SK, Mackinnon JM, Roberts MJ, Buchanan CJ, Kruczkiewicz P, Jokinen CC, Thomas JE, Nash JH, Gannon VP, Marshall B, Pollari F, Clark CG: Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. J Clin Microbiol 2012;50:788–797.

22 Eisenstein M: Oxford nanopore announcement sets sequencing sector abuzz. Nat Biotechnol 2012;30:295–296.