Whole genome sequencing for foodborne disease surveillance

Landscape paper

World Health Organization

Whole genome sequencing for foodborne disease surveillance

Landscape paper

World Health Organization

Whole genome sequencing for foodborne disease surveillance: landscape paper

ISBN 978-92-4-151386-9

Suggested citation. Whole genome sequencing for foodborne disease surveillance: landscape paper. Geneva: World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Printed in Switzerland

Contents

Acronyms and abbreviations

AMR	antimicrobial resistance
API	application program interface
CFSAN	Center for Food Safety and Applied Nutrition (USA)
CGE	Center for Genomic Epidemiology, Denmark Technical University (Denmark)
CLI	command line interface
EUCAST	European Committee on Antimicrobial Susceptibility Testing
FAO	Food and Agriculture Organization of the United Nations
FBD	foodborne disease
FDA	Food and Drug Administration (USA)
Gb	gigabase
GO	gene ontology
GUI	graphical user interface
HUS	haemolytic uraemic syndrome
IHR (2005)	International Health Regulations (2005)
INSDC	International Nucleotide Sequence Database Collaboration
IRIDA	Integrated Rapid Infectious Disease Analysis
IT	information technology
kb	kilobase
LIMS	laboratory information management system
MLST	multilocus sequence typing
MLVA	multilocus variable-number tandem-repeat analysis
NCBI	National Center for Biotechnology Information (USA)
NGS	next generation sequencing
OIE	World Organisation for Animal Health
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PFGE	pulsed-field gel electrophoresis
QA	quality assurance
QC	quality control
SBL	sequencing-by-ligation
SBS	sequencing-by-synthesis
SENASICA	Servicio Nacional de Sanidad, Inocuidad y Calidad Agroalimentaria
SMRT	single-molecule real-time sequencing
SNP	single nucleotide polymorphism
SNVPhyl	single nucleotide variant phylogenomics
SRA	sequence read archive
STEC	Shiga-toxin-producing Escherichia coli
Stx	Shiga toxin
WGS	whole genome sequencing
WHO	World Health Organization

Acknowledgements

The World Health Organization (WHO) expresses sincere thanks to all the authors and other reviewers of this paper.

Contributing authors

David Aanensen, Imperial College London, London, England; Clara Amid, European Bioinformatics Institute European Molecular Biology Laboratory, Cambridge, England; Stephen Baker, Oxford University Clinical Research Unit, Ho Chi Minh, Viet Nam; Claudio Bandi, Università degli Studi di Milano, Milan, Italy; Eric W. Brown, United States Food and Drug Administration (FDA), Silver Spring, MD, United States of America (USA); Josefina Campos, Instituto Nacional de Enfermedades Infecciosas, Buenos Aires, Argentina; Guy Cochrane, European Bioinformatics Institute European Molecular Biology Laboratory, Cambridge, England; Francesco Comandatore, Università degli Studi di Milano, Milan, Italy; Tim Dallman, Public Health England, London, England; Xiangyu Deng, University of Georgia, Griffin, GA, USA; Gordon Dougan, Department of Medicine, University of Cambridge, Cambridge, England; Rita Finley, Public Health Agency of Canada, Guelph, Canada; Alejandra García Molina, SENASICA, Mexico City, Mexico; Peter Gerner-Smidt, Centers for Disease Control and Prevention, Atlanta, GA, USA; Sara Goodwin, Cold Spring Harbor, NY, USA; Tine Hald, Technical University of Denmark, Copenhagen, Denmark; Zhaila Isaura Santana Hernández, Servicio Nacional de Sanidad, Inocuidad y Calidad Agroalimentaria (SENASICA), Mexico City, Mexico; Kirsty Hope, New South Wales Ministry of Health, Canberra, Australia; William Hsiao, British Columbia Centre for Disease Control Public Health Laboratory, Vancouver, Canada; Claire Jenkins, Public Health England, London, England; Katherine Littler, Wellcome Trust, London, England; Ole Lund, Technical University of Denmark, Copenhagen, Denmark; Megge Miller, South Australian Department for Health and Ageing, Adelaide, Australia; Alejandra García Molina, SENASICA, Mexico City, Mexico; Jacob Moran-Gilad, Ben Gurion University of the Negev, Beer-Sheva and Israeli Ministry of Health, Jerusalem, Israel; Nicola Mulder, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa; Celine Nadon, Public Health Agency of Canada, Ottawa, Canada; Eric Ng'eno, Kenya Medical Research Institute, Nairobi, Kenya; Collins Owuor, Kenya Medical Research Institute and Wellcome Trust Research Institute, Kilifi, Kenya; Julian Parkhill, Wellcome Trust Sanger Institute, Cambridge, England; Mirko Rossi, University of Helsinki, Helsinki, Finland; Jørgen Schlundt, Nanyang Technological University, Singapore; Vitali Sintchenko, University of Sydney, Sydney, Australia; Nicholas R. Thomson, Wellcome Trust Sanger Institute and London School of Hygiene and Tropical Medicine, London, England.

Scientific editors

Amy Cawthorne, WHO, Geneva, Switzerland; Eelco Franz, National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands; Rene Hendrikssen, Technical University of Denmark,

Copenhagen, Denmark; Simone Magnino, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna, Pavia, Italy; Stefano Pongolini, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna, Parma, Italy; and Eric Stevens, FDA, Silver Spring, MD, USA.

Introduction

The most effective way to manage foodborne disease (FBD) threats is to detect them rapidly, understand them and respond to them. To do this, public health authorities around the world need surveillance and response systems capable of:

- rapidly detecting food safety events and FBD outbreaks;
- monitoring trends in priority FBDs, in order to assess the circulation and variation of human pathogens, including monitoring antimicrobial resistance (AMR) patterns.

Whole genome sequencing (WGS) provides the highest possible microbial subtyping resolution available to public health authorities for the surveillance of and response to FBDs. Used as part of a surveillance and response system, it has the power to increase the speed with which threats are detected and the detail in which the threats are understood, and ultimately lead to quicker and more targeted interventions. Given its power, all countries are encouraged to explore how the technology can be used to improve their surveillance and response systems. While this landscaping paper and the accompanying country decision-making tool described below are focused primarily on FBDs, the discussion and advice are pertinent to all infectious disease surveillance and response systems.

To help countries understand the implications, costs and benefits of investing in WGS as a tool to strengthen national surveillance and response systems, WHO convened a meeting in Washington, DC, on 10–13 January 2017. The meeting brought together technical experts from around the world to discuss how WGS could be used in developing countries to support FBD surveillance and response. On the basis of the meeting, WHO will produce a guidance document to support countries wishing to use WGS to strengthen FBD surveillance and response.

To ensure that the guidance is comprehensive and relevant, this landscape paper has been prepared by technical experts from laboratories and public health authorities. It summarizes some of the benefits and challenges inherent in the implementation of WGS and describes some of the issues developing countries may face. It also provides an evidence base for some of the approaches to be considered in the guidance document.

This landscape paper aims to:

- describe the public health impact of WGS as a tool for strengthening integrated surveillance along the food chain, with a specific focus on its application to FBD surveillance;
- identify the barriers to implementation in low- and middle-income countries;
- summarize the current state of WGS technology; and
- describe how different people working in public health use information from WGS.

1. Whole genome sequencing: the future of FBD surveillance and outbreak response

WGS provides much greater strain discrimination than other methods for typing foodborne bacterial pathogens. In addition, it provides an all-in-one test in the sense that information usually obtained from other typing methods (including serotyping, molecular subtyping and resistance profiling) can be extracted *in silico* from the sequence data. WGS-derived phylogenetic analysis improves cluster resolution and is an invaluable tool in epidemiological investigations. Retrospective studies have demonstrated the utility of WGS for detection of FBD outbreaks, case definitions and case ascertainment *(1-7)*. Recently, a number of national public health bodies have used WGS for real-time surveillance of foodborne bacterial pathogens *(8-13)*.

1.1 Public health surveillance

1.1.1 Subtyping of pathogens for surveillance and outbreak investigation

WGS offers high-resolution subtyping of different bacterial, viral, fungal and parasitic pathogens *(14-19)*. This capability can be used for retrospective comparison of microorganisms associated with epidemiologically suspected outbreaks or for prospective laboratory surveillance of high-burden diseases, such as listeriosis and salmonellosis *(20)*.

Retrospective comparisons to test epidemiological hypotheses are generally guided by public health professionals and usually involve epidemiologists, (molecular) microbiologists, bioinformaticians, and clinicians. The value of such comparisons increases when they are performed within the time period of the outbreak investigation (e.g. the definition of an outbreak case includes WGS confirmation) and when the investigation involves cases from neighbouring jurisdictions and several laboratories.

In contrast, WGS-based prospective surveillance relies on monitoring of cases by jurisdictional public health laboratories; alerts are generated when clusters of pathogens with similar genomes are identified in a limited geographical area or time period. This prospective surveillance places greater responsibility on public health laboratories and requires close collaboration between the laboratory and public health units, including real-time data sharing arrangements. WGS-based surveillance often allows cases that are misclassified by other laboratory methods, including other molecular subtyping methods, to be implicated or ruled out of an outbreak. Genomics-based surveillance relies on the assumption that pathogens with similar genomes (e.g. few SNPs between them) come from a common source. While this is feasible, particularly with phylogenetic methods that enable closely related isolates to be clustered together, it is the combination of relevant epidemiological information that allows the transmission pathways to be inferred in detail. The continuous synthesis and evaluation of genomic and epidemiological evidence should be encouraged to increase the usefulness of the data *(20-22)*.

1.1.2 Comparison of WGS with traditional methods for real-time surveillance

Previously, the standard typing methods for many foodborne bacterial pathogens were antigen testing, pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem-repeat analysis (MLVA). However, these methods have a lower level of strain discrimination than WGS. A study by den Bakker et al. *(23)* clearly demonstrated increased resolution of whole genome cluster analysis, and subsequent outbreak detection in common PFGE pattern types of *Salmonella* Enteritidis. They showed that using WGS for real-time surveillance facilitated the detection of numerous potential clusters that would have gone undetected by PFGE. During this study, WGS also detected more cases associated with known outbreaks. In one example, additional clinical isolates from patients in surrounding communities, not previously associated with a specific outbreak, expanded the number of possible outbreak cases from seven to 16. Knowledge of these cases at the time of the outbreak may have improved the chances of finding the outbreak source, which was never resolved.

For *Listeria monocytogenes*, Kwong et al. *(12)* found that WGS offered higher resolution than PFGE and serotyping for isolates; they were able to use this to infer the likely mode of transmission or point-source exposure in outbreaks of listeriosis.

Dallman et al. *(9)* compared WGS and MLVA for Shiga toxin-producing *Escherichia coli* (STEC) O157. They found no significant difference between the two methods in timeliness of cluster detection. However, the time to cluster completion (when all cases of a cluster have been identified) from the initial cluster event was significantly higher with WGS than with MLVA. For example, during an outbreak associated with drinking raw milk, real-time MLVA surveillance identified an additional nine isolates that appeared to be closely related to the outbreak *(24)*. However, it was uncertain whether these additional cases (who did not initially report consumption of raw milk) were linked to the outbreak. In contrast, WGS confirmed that four of the nine additional cases were from the same outbreak. Subsequently, in-depth epidemiological investigations, driven by the forensic certainty of the WGS analysis, provided evidence that these additional four cases had consumed raw milk from the implicated farm; there was no evidence of consumption of raw milk by the remaining five cases.

1.1.3 WGS detects outbreaks taking place under the surveillance radar

In a study in the USA, den Bakker et al. *(23)* observed a small cluster of isolates of *S.* Enteritidis obtained over a 2.5-year period, suggesting a persistent point source in the environment. Kanagarajah et al. also uncovered a previously undetected outbreak of *S.* Enteritidis phage type (PT8) linked to handling of reptile feeder mice, or snakes infected by the mice, that had been ongoing in the United Kingdom for four years *(25)*. The number of cases each month was relatively low compared with other cases of *S.* Enteritidis PT8, and epidemiological links to handling of reptiles were confounded by unlinked cases of S. Enteritidis PT8 that were not differentiated from the outbreak cluster by phage typing. This outbreak demonstrated the potential of WGS to identify low-level, continuous-transmission outbreaks that previously went undetected. The unprecedented level of strain differentiation that WGS affords allows a very specific case definition that facilitates highly accurate case ascertainment and a robust, focused epidemiological investigation. SNP

typing of the core genome also provided evolutionary context, making it possible to confidently link cases from four years earlier to the contemporary cluster.

1.2 Additional information from phylogenetic analysis

1.2.1 Outbreak investigation and source-finding

Because mutational drift is sequential, phylogenetic methods can be used to study variation in genomes and determine evolutionary relationships. It is therefore possible to explore the deeper phylogenetic relationships between strains in order to uncover clues to the origin of an outbreak strain or to determine the most likely mode of transmission. For example, whole genome cluster analysis by den Bakker et al. *(23)* revealed that an outbreak in a long-term care facility belonged to the same monophyletic lineage as isolates associated with a previous outbreak caused by contaminated shelled eggs, indicating that shelled eggs were likely to be a common source of infection. Additionally, Hoffmann et al. *(10)* demonstrated the power of combining genomic information with an isolate's geographical origin in a foodborne outbreak involving *Salmonella* Bareilly. The retrospective study highlighted how WGS data and epidemiological information could provide immediate clues to the source of contamination, even if it is halfway around the world.

During an outbreak of STEC O157 in Northern Ireland, strains held in the Public Health England (PHE) WGS database that clustered most closely with the outbreak strains were associated with foreign travel to Egypt and Israel *(26)*. Although the contaminated food source was never confirmed by microbiological testing, epidemiological analysis implicated dried parsley imported from the Mediterranean region as the most likely vehicle. In 2013 in the United Kingdom, two concurrent outbreaks of Shiga-toxin-producing *Escherichia coli* O157 were linked with the consumption of watercress. Analyses of sporadic isolates obtained during routine surveillance indicated that in one outbreak the contaminated watercress was domestically produced (this was later confirmed by environmental testing) and that the watercress linked to the other outbreak was probably imported *(5)*. Similarly, a large-scale international outbreak of *Salmonella* Enteritidis was investigated using WGS. An important finding was that isolates with multiple MLVA-types can be genomically similar and considered part of the outbreak. Importantly, WGS can also be used to exclude outbreak relations. For example, recently a parallel increase in S. Newport infections was notified in The Netherlands and Ireland but WGS revealed these were not related and there was no cross-border outbreak.

1.2.2 Source attribution

The main goal of source attribution analysis is to partition human disease (for example, salmonellosis or campylobacteriosis) over a number of putative sources of infection. Quantitative estimates of the relative contributions of different sources to human disease is crucial in setting priorities for public health interventions and measuring the impact of such interventions. A detailed overview of definitions, terminology, and methodologies for source attribution has been presented elsewhere *(27-29)*. So far, efforts to quantify the relative contribution of different (animal, food, and environmental) sources to human illness have mostly relied on phenotypic (e.g. serotyping, antimicrobial resistance profiling, etc.) and genotypic subtyping methods other than WGS (e.g. multilocus sequence typing (MLST), MLVA). Given

the source specificity of certain pathogen subtypes and assuming a unidirectional transmission pathway, from sources to humans (with humans representing the endpoint), the relative contribution of each source to human cases can be inferred probabilistically by comparing the human and source subtype distributions *(30)*. Defining the optimal level of discrimination for source attribution is a challenge and depends partly on the level of clonality and degree of host association of the pathogen under investigation. Ideally, source attribution methods should allow for some genetic diversity between isolates, but only to the extent that they can still be assumed to originate from the same source *(31)*. Because of its superior resolution, WGS has the potential to significantly improve source attribution models. However, the use of WGS in source attribution requires the development of new modelling approaches that can handle the large amount of data that is generated, such as population genetics models. In addition, an optimum level of resolution should be defined to tune the high discriminatory power of WGS data according to the specific pathogen in question.

1.3 Predicting emerging threats

Analyses of surveillance data can monitor the emergence of virulent clones within a population of foodborne bacterial pathogens. Infection with STEC O157 can progress to haemolytic uraemic syndrome (HUS), and there is a significant association between the development of HUS and cases infected with STEC O157 harbouring the Shiga toxin subtype, Stx2a. The acquisition of the Stx2a subtype in STEC O157 occurred relatively recently and is likely to explain the recent emergence of STEC O157 as a clinically significant pathogen *(9)*. The Stx2a-encoding phage has been acquired by STEC O157 on multiple occasions, highlighting the potential for new, highly virulent clones to emerge. Of concern is that once Stx2a is integrated in a population it tends to be maintained. Such analyses can provide insight into the dynamics of STEC O157 transmission on a national and international scale. For example, deeper phylogenetic analysis of a strain of STEC O157 associated with raw milk identified a highly pathogenic clade of STEC O157 PT21/28 harbouring Stx2a only *(24)*. Use of WGS for real-time public health surveillance of foodborne pathogens enables us to monitor the emergence of pathogenic variants and the associated modes of transmission *(32, 33)*.

1.4 Monitoring antimicrobial resistance in foodborne pathogens

Antimicrobial resistance is a growing global public health problem linked to the increased use of both human and veterinary antimicrobial drugs. With the recent focus on a "One Health" approach that links agricultural, animal, and human health, there is clear evidence of the public health risk posed by animal reservoirs for the transmission of antimicrobial-resistant strains of zoonotic bacteria to humans. This transmission pathway includes the food chain, and monitoring AMR in foodborne pathogens isolated from clinical (i.e. human and animal), food and environmental samples may help to understand and mitigate the public health risk of the transmission of resistant strains from animals to humans *(34)*.

Microbiologists can reliably detect drug susceptibility and resistance from genome sequences for many bacterial and viral pathogens with established catalogues of molecular markers of drug resistance *(16,*

19, 35, 36). WGS as a frontline method for public health protection will improve antimicrobial resistance profiling and biological risk prediction *(21, 37)*.

The implementation of WGS for real-time public health surveillance of foodborne pathogens facilitates the assessment of AMR by identifying the complement of antibiotic resistance genes in an organism. Several studies examining foodborne and other pathogens have shown a high degree of correlation between clinical resistance and the presence of acquired resistance genes for most drug classes *(37, 38)*. Recently, the European Committee on Antimicrobial Susceptibility Testing *(39)* concluded that available published evidence does not currently support use of WGS-inferred susceptibility alone as a guide in clinical decision-making. However, the Committee acknowledged that this approach may replace phenotypic testing for surveillance purposes in the near future *(39)*.

1.5 References

1. Angelo KM, Chu A, Anand M, Nguyen TA, Bottichio L, Wise M, et al. Outbreak of Salmonella Newport infections linked to cucumbers – --United States, 2014. MMWR Morb Mortal Wkly Rep. 2015;64(6):144-7.
2. Ashton PM, Peters T, Ameh L, McAleer R, Petrie S, Nair S, et al. Whole Ggenome Ssequencing for the Rretrospective Iinvestigation of an Ooutbreak of Salmonella Typhimurium DT 8. PLoS Curr. 2015;7.
3. Byrne L, Fisher I, Peters T, Mather A, Thomson N, Rosner B et al. A multi-country outbreak of Salmonella Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012. Euro Surveill. 2014;19(31):6-13.
4. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T et al. Tracing origins of the Salmonella Bareilly strain causing a food-borne outbreak in the United States. J Infect Dis. 2016;213(4):502-8.
5. Jenkins C, Dallman TJ, Launders N, Willis C, Byrne L, Jorgensen F et al. Public health investigation of two outbreaks of Shiga toxin-producing Escherichia coli O157 associated with consumption of watercress. Appl Environ Microbiol. 2015;81(12):3946-52.
6. Kvistholm Jensen A, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S et al. Whole-genome sequencing used to investigate a nationwide outbreak of listeriosis caused by ready-to-eat delicatessen meat, Denmark, 2014. Clin Infect Dis. 2016;63(1):64-70.
7. Awofisayo-Okuyelu A, Arunachalam N, Dallman T, Grant KA, Aird H, McLauchlin J et al.. An outbreak of human listeriosis in England between 2010 and 2012 associated with the consumption of pork pies. J Food Prot. 2016;79(5):732-40.
8. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol. 2016;54(8):1975-83.
9. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing Escherichia coli O157. Clin Infect Dis. 2015;61(3):305-12.
10. Gardy JL, Loman NJ, Rambaut A. Real-time digital pathogen surveillance—the time is now. Genome Biol. 2015;16:155.
11. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis. 2016;63(3):380-6.
12. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM et al. Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. J Clin Microbiol. 2016;54(2):333-42.
13. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ. 2016;e1752.
14. European Centre for Disease Prevention and Control. Expert opinion on whole genome sequencing for public health surveillance. Stockholm: ECDC; 2016.
15. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. N Engl J Med. 2013;369(13):1195-205.
16. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469-74.

17. Outhred AC, JP, Suliman B, Hill-Cawthorne GA, Crawford ABH, Marais BJ, Sintchenko V. Added value of whole-genome sequencing for management of highly drug-resistant tuberculosis. J Antimicrob Chemother. 2015;70(4):1198-202.
18. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N et al. Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol. 2013;51:232-237.
19. Westblade LF, van Belkum A, Grundhoff A, Weinstock GM, Pamer EG, Pallen MJ et al. Role of clinicogenomics in infectious disease diagnostics and public health microbiology. J Clin Microbiol. 2016;54(7):1686-93.
20. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ. 2015;350:h1314.
21. European Centre for Disease Prevention and Control. Technical Report: ECDC roadmap for integration of molecular typing into European-level surveillance and epidemic preparedness. Version 2.1, 2016-2019. Stockholm: ECDC; 2016.
22. Gonzalez-Candelas F, Bracho MA, Wrobel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol. 2013;11:76.
23. den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z et al. Rapid whole-genome sequencing for surveillance of Salmonella enterica serovar enteritidis. Emerg Infect Dis. 2014;20(8):1306-14.
24. Butcher H, Elson R, Chattaway MA, Featherstone CA, Willis C, Jorgensen F et al. Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk. Epidemiol Infect. 2016;144(13):2812-23.
25. Kanagarajah S, Waldram A, Dolan G, Jenkins C, Ashton PM, Carrion Martin AI et al. Whole genome sequencing reveals an outbreak of Salmonella Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. J Food Microbiol. 2017.
26. Lauren A, Cowley TJD, Fitzgerald S, Irvine N, Rooney PJ, McAteer SP et al. Short-term evolution of Shiga toxin-producing Escherichia coli O157:H7 between two food-borne outbreaks. Microbial Genomics. 2016.
27. Pires SM. Assessing the applicability of currently available methods for attributing foodborne disease to sources, including food and food commodities. Foodborne Pathog Dis. 2013; 10(3):206-13.
28. Pires SM, Evers EG, van Pelt W, Ayers T, Scallan E, Angulo FJ, et al., Attributing the human disease burden of foodborne infections to specific sources. Foodborne Pathog Dis. 2009; 6(4):417-24.
29. Pires SM, Vieira AR, Hald T and Cole D. Source attribution of human salmonellosis: An overview of methods and estimates. Foodborne Pathog Dis. 2014; 11(9):667-76.
30. Mughini-Gras L, Franz E and van Pelt W. New paradigms for salmonella source attribution based on microbial subtyping. Food Microbiol. 2018; 71:60-67.
31. Franz E, Gras LM, GL, Dallman T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Current Opinion in Food Science. 2016;8:74-9.
32. Bielaszewska M, Mellmann A, Bletz S, Zhang W, Kock R, Kossow A et al. Enterohemorrhagic Escherichia coli O26:H11/H-: a new virulent clone emerges in Europe. Clin Infect Dis. 2013;56(10):1373-81.
33. Petrovska L, Mather AE, AbuOun M, Branchu P, Harris SR, Connor T et al. Microevolution of monophasic Salmonella Typhimurium during epidemic, United Kingdom, 2005-2010. Emerg Infect Dis. 2016;22(4):617-24.
34. Day M, Doumith M, Jenkins C, Dallman TJ, Hopkins KL, Elson R et al. Antimicrobial resistance in Shiga toxin-producing Escherichia coli serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. J Antimicrob Chemother. 2017;72(1):145-152.
35. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med. 2014;6(254):254ra126.
36. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med. 2012;4(148):148ra16.
37. Tyson GH, McDermott PF, Li C, Chen Y, Tadesse DA, Mukherjee S et al. WGS accurately predicts antimicrobial resistance in Escherichia coli. J Antimicrob Chemother. 2015;70(10):2763-9.
38. Tyson GH, Zhao S, Li C, Ayers S, Sabo JL, Lam C et al. Establishing genotypic cutoff values to measure antimicrobial resistance in Salmonella. Antimicrob Agents Chemother. 2017;61(3).
39. European Society of Clinical MIcrobiology and Infectious DIseases. Report from the EUCAST SUbcommittee on the role of Ehole Genome Sequencing (WGS) in antimicrobial susceptibility testing of bacteria. (http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Consultation/2016/EUCAST_WGS_report_consultation_20160511.pdf, accessed 21 March 2018).

2. WGS as a tool to strengthen integrated surveillance

2.1 Overview of integrated foodborne disease surveillance

FBD surveillance aims to reduce the burden of illness caused by eating contaminated foods. The objectives of surveillance include monitoring of disease trends, estimation of disease burden, identification of vulnerable groups, determination of sources of contamination and routes of transmission, and identification and control of outbreaks *(1)*. The output from the surveillance system is used to inform policies and improve prevention strategies. The levels of surveillance effort range from no formal surveillance, to syndromic surveillance, to laboratory-based surveillance and finally integrated food chain surveillance; each level requires increased infrastructure and resources *(2)*.

Integrated FBD surveillance combines data from different parts of the food chain (farm to fork) to provide comprehensive information for identifying and confirming outbreaks, monitoring disease trends, identifying risk factors and populations, and improving food production and public health practices. The goals and objectives of an integrated approach are to identify sources and patterns of endemic and emerging disease, and to support an efficient and coordinated multi-agency response to health risks along the food chain *(3)*. Pathogenic microorganisms can enter the food chain at any point in the farm-to-fork continuum (e.g. livestock feed, farm production site, slaughterhouse, packing plant, manufacture, retail, and home preparation). Integrated FBD surveillance combines data from multiple sources, including the environment, farms, processing plants, retail outlets, and hospitals. The routine collection, collation and interpretation of all these data greatly improve the ability to rapidly trace sources of contamination and estimate the relative contribution of different food sources to human FBD *(4)*.

Integrated food chain surveillance is resource-intensive and is often implemented only in high-income countries and for specific pathogens *(3, 5, 6)*. It requires: an adequate health care and regulatory infrastructure to support the collection and processing of food, clinical and laboratory samples and data; high quality, dedicated laboratory and epidemiological personnel to analyze them; and increasingly, specialized data scientists to integrate and facilitate sharing of cross-sectoral heterogeneous data. Standardization of laboratory methods and cross-sectoral reporting are critical if successful data integration is to be achieved. In the past, the use of different laboratory techniques, including microbiological typing techniques, often made the collation and comparison of data impossible. With the advent of WGS, all sectors will be able to generate compatible data from multiple sources, making collaboration a real possibility. This technology has the potential to improve the level of integration of disease surveillance across the food chain and the health system, and ultimately to encourage leaps forward in the development of safe food chains.

2.2 High accuracy matching of pathogens across the animal, food, environmental and human sectors

The establishment of PulseNet in various parts of the world provided an opportunity to introduce, in a standardized manner, a new molecular method that, at the time, provided greater granularity for identifying

links among human cases and between human and non-human samples. This success was made possible through the development of national databases of PFGE patterns generated by public health, veterinary and food laboratories; these were then evaluated, assessed and interpreted in an integrated manner. As a result, PFGE became critical in the identification of clusters, leading to an increase in the number of outbreaks investigated and providing better evidence of links between human cases and the sources of infection.

The introduction of WGS has increased the ability to distinguish between outbreak-related and sporadic cases, to link sporadic cases to particular food and animal sources, and to identify points of contamination and areas requiring intervention during product trace-back and recalls. At the same time, WGS can be used to exclude wrongly suspected sources of infection (e.g. specific food commodities), which prevents economic damage that was previously unavoidable with lower-resolution methods *(7)*.

This greater sensitivity will allow limited resources to be better focused on outbreaks that are more likely to be solved, which could lead to faster resolution of outbreak investigations; however, it is important to note that WGS will identify more links and clusters than are able to be acted upon. In addition, through the timely and routine integration of food and animal sequence information, links between these and human isolates can be identified. This has been done successfully in Europe in an investigation of a multicountry *Salmonella* Enteritidis outbreak associated with eggs in 2014 *(8)*. Through the use of WGS on *S.* Enteritidis isolates from humans and eggs from various countries, investigators were able not only to link illnesses to contaminated eggs, but also to trace back the strain to a specific German company, allowing control measures to be implemented. Similarly, in the United States of America, WGS has been used routinely since 2013 to detect and investigate cases of *Listeria monocytogenes*, including the testing of food and environmental samples *(9)*. The use of WGS on *Listeria* strains has strengthened the links between human, food and environmental isolates. This has resulted in more accurate detection of clusters, avoided investigations of outbreaks not deemed as "true outbreaks", and allowed more outbreaks to be successfully resolved: nine investigations were successfully concluded in the second year of using WGS compared with two in the year before WGS was implemented. To ensure that WGS is as effective as possible, data from across the food continuum (animal, food, environment and humans) should be publicly shared and distributed, so that sequence data can be analysed in an integrated manner. This is very much how things are done through PulseNet and other databases (e.g. GenomeTrakr) that are linked to increase integration among various stakeholders.

It is widely acknowledged that WGS needs to be introduced gradually, not only in public health but across all sectors. The implementation of WGS will require extra resources and technical training, which will be acquired over time. It is also important to ensure that historical information is not lost *(10)* and that trends among humans, animals and food can be monitored during the transition period. Sequencing a representative sample of historical isolates can contribute to maintaining continuity in monitoring of pathogen populations. In addition, for certain currently used genotyping methods, typing profiles can be reproduced *in silico* from WGS data of isolates, thus ensuring comparability of results between current methods and WGS. This is already the case for the extensively used MLST and MLVA. Furthermore, introduction of WGS requires the method to be properly validated and interpretation criteria developed, allowing a standard

approach to be achieved and agreed upon by the different actors. Sharing this development and validation process not only supports the scientific validity of the developed methodology and interpretation criteria but also ensures that the information produced is widely accepted and that it continues to be useful for surveillance, source attribution, outbreak detection, and trace-back investigations *(11)*. In this process, it is particularly critical to define the interpretation criteria. They are the values of WGS outputs (e.g. the number of genetic loci varying between two genomes or the topology of a phylogenetic tree) that support or exclude the assignment of isolates to an outbreak or inform about the evolution of a pathogen, etc. Such interpretation criteria should be extrapolated through the application of WGS to real-life cases of infection for which all the significant epidemiological details are available and represent the known variables of the study. So, for instance, the investigation of a set of human, food and animal isolates known to belong to epidemiologically demonstrated outbreaks will provide information about the actual number of intra-outbreak genetic differences between isolates. This number and its variability across a variety of similar outbreaks will represent a criterion for the interpretation of future unknown field scenarios. This process of criteria identification and validation is evidently a very cross-sectoral one, in which epidemiologists and experts from public health, food hygiene and animal health must put together high quality information and share the interpretation of analysis outcomes.

2.3 Coordinating the use of WGS across public health, food safety and regulatory agencies

Setting up an integrated surveillance mechanism that builds on the strengths of WGS requires a coordinated approach across the "One Health" spectrum, including public health, food safety, veterinary health and regulatory agencies. Achieving such coordination is challenging, because of differences in organizational aspects of the involved bodies and agencies, local or regional cultural and political factors, technical and operational considerations, and scientific aspects (Figure 2.1).

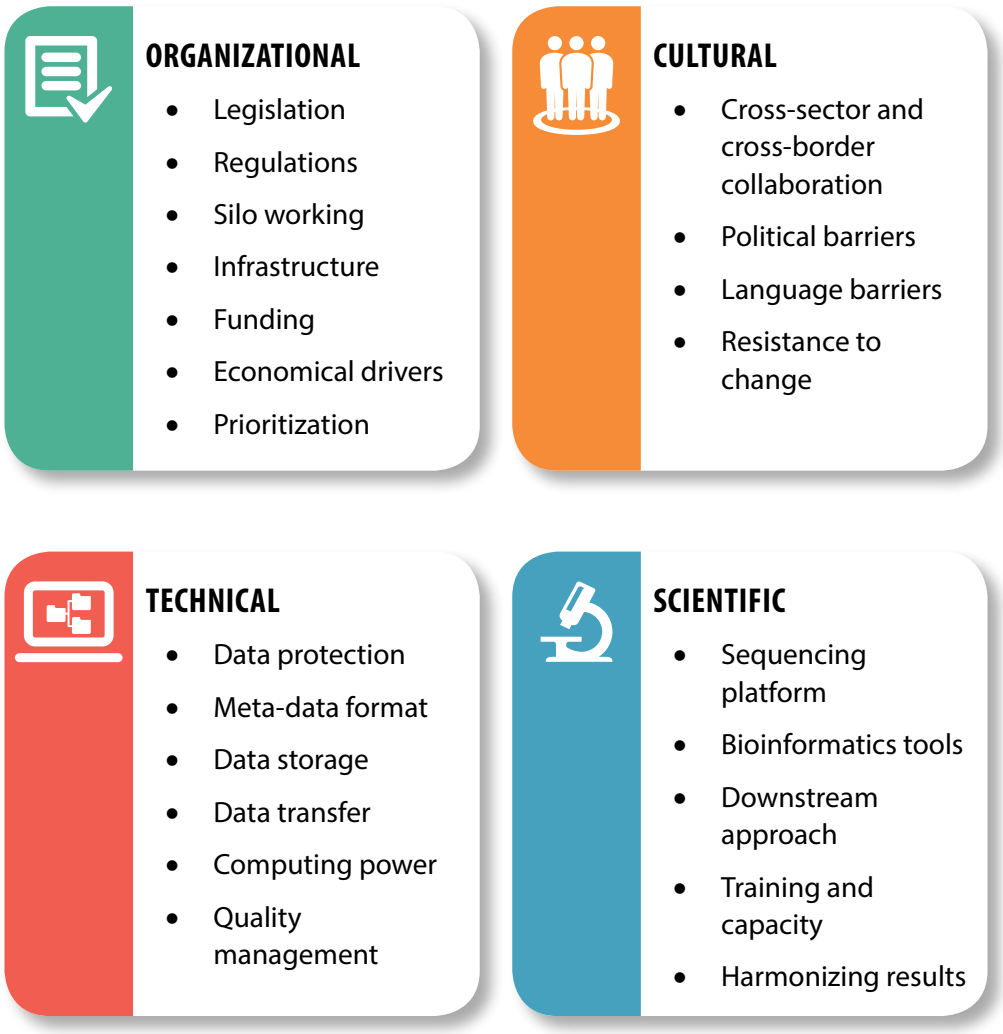
2.3.1 Organizational and cultural aspects

From an organizational perspective, communication between different agencies across the One Health spectrum and engagement of relevant stakeholders are key to ensuring a coordinated approach in any integrated food chain surveillance programme. Existing legislation and regulations at local and regional levels need to be taken into account, and possibly modified, to facilitate the penetration of WGS-based approaches. Multiple priorities and requirements set by different actors could produce varying approaches to surveillance; ensuring transparency from the beginning will be important when implementing WGS.

Different sectors may be ready to move ahead with WGS at different times. Readiness will be influenced by many factors, including the availability of infrastructure (especially sequencing technology and computational and bioinformatics capabilities), internal and external funding and related economic drivers (e.g. import and export of food or animals). No less important are the cultural and political aspects (and corresponding barriers) of working across the different sectors, and commonly across international borders. Information needs to be shared, although not necessarily routinely. Relevant sequence and meta data should be uploaded to international databases in order to link food to human illness at the global

FIGURE 2.1

Challenges of coordinating WGS for integrated food chain surveillance



level. The support and buy-in of the technical and scientific communities are needed to move away from traditional methods and adopt the new technology.

2.3.2 Technical and scientific aspects

Challenges emerge with the increasing diversity of sequencing platforms and especially downstream analysis tools. There are extreme difficulties in standardisation of bioinformatics analysis, even in the most modern settings *(12)*. Ensuring the quality and robustness of sequencing and downstream analysis is becoming a major challenge for wide adoption of WGS-based surveillance *(13)*. Even if different tools or computational approaches are being used, it is crucial to harmonise their outputs in order to create meaningful data that would inform public health actions *(14)*. Therefore, the implementation of WGS in low resource settings will mandate robust and automated downstream solutions *(15)*, especially in light of the limitations in sequencing and computational infrastructures and paucity of skilled and trained personnel.

The introduction of technology to such settings should thus be coupled with appropriate capacity building efforts *(15)*, for both the 'wet' and 'dry' laboratory parts.

Ensuring that sufficient metadata are available across all sectors to provide context to the WGS results is central to making most effective use of this technology. The data should include information on: cases of illness (time, place, person), implicated foods (source, time and place), implicated animals (source, time and place) and related environments (source, time and place). These metadata should complement the wealth of data generated by traditional and molecular microbiology laboratories involved in the process, including clinical microbiology laboratories, public health and reference laboratories, and food, water and environmental laboratories, as well as *in silico* by analysis of WGS data. The latter includes species calling, inference of antimicrobial resistance and virulence, epidemiological phylotyping, and – in the near future – source attribution or microbial risk assessment. Microbiological results and metadata should be collected, stored, analysed and reported in a standardized manner, to facilitate the mining, transfer and sharing of the data between sectors.

In low resource setting, data storage and transfer could also pose a major challenge due to infrastructure and technical requirements. Opportunities associated with emerging sequencing technologies which would enable selective and portable / deployable sequencing in a low-resource environment without the need for significant capital investments may be around the corner *(16)*. But even portable sequencing technologies (e.g. Nanopore), require readily available analysis platforms. If raw data can be processed close to the source of sequencing, then the results can be reported without the need to transfer large amount of data. This is currently feasible for pathogen identification but due to technical limitations of current deployable sequencers, detailed source tracking or variant analysis (based on SNP or wgMLST) analysis for genotyping purpose is not yet feasible. In such cases, a two-staged approach may be appropriate, where fast identification can be made at the low resource setting and shared quickly whereas the more detailed epidemiological analysis requiring detailed genotyping analysis of the raw data can be performed at a later stage, following data upload. Crowd sourcing and cloud computing may also provide future solutions to such challenges *(16)*.

An additional challenge might be the introduction of culture-independent diagnostic testing or molecular tools in the different sectors. While molecular diagnostics may improve sensitivity and specificity, increasing reliance on these methods may result in a loss of critical information, as pathogens are less often recovered by culture and submitted for analysis, thus hampering surveillance efforts. Isolation of foodborne pathogens by culture therefore remains a critical step. Looking forward, further complexity could be envisaged and should be mitigated, as culture-independent approaches based on shotgun metagenomics of clinical and environmental samples are developed and implemented.

2.4 References

1. Ford L, Miller M, Cawthorne A, Fearnley E, Kirk M. Approaches to the surveillance of foodborne disease: a review of the evidence. Foodborne Pathog Dis. 2015;12(12):927-36.
2. Methods for foodborne disease surveillance in selected sites: report of a WHO consultation. Geneva: World Health Organization; 2002 (WHO/CDS/CSR/EPH/2002.22; 2002).
3. GalanisE, Parmley J, De WithN, & Group, B. C. I. S. O. F. P. W. Integrated surveillance of Salmonella along the food chain using existing data and resources in British Columbia, Canada. FRIN, 2012. 45: p. 795-801.
4. PiresSM, Vieira AR,Hald T, Cole D. Source attribution of human salmonellosis: an overview of methods and estimates. Foodborne Pathog Dis, 2014;11(9):667-76.
5. David JM, Danan C, Chauvin C, Chazel M, Souillard R, Brisabois A, et al. Structure of the French farm-to-table surveillance system for Salmonella. Revue Med Vet 2011;162:489-500.
6. Hald T, Wegener HC, Borck B, Wong DLF, Baggesen DL, Madsen M, et al. The Integrated Surveillance of Salmonella in Denmark and the Effect on Public Health. in JM Smulders & JD Collins (eds), Risk management strategies: monitoring and surveillance. Food safety assurance and veterinary public health. Wageningen: Wageningen Academic Publisher. 2004; 213-238.
7. Deng X, den Bakker HC, Hendriksen RS. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. Ann Rev Food Sci Technol. 2016;7:353-74.
8. Inns T, Lane C, Peters T, Dallman T, Chatt C, McFarland N et al. A multi-country Salmonella Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. Euro Surveill. 2015;20(16).
9. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis. 2016;63(3):380-6.
10. Sabat AJ, Budimir A, Nashev D, Sa-Leao R, van Dijl J, Laurent F et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013;18(4):20380.
11. Use of whole-genome sequencing (WGS) of food-borne pathogens for public health protection. EFSA Scientific Colloquium Summary Report 20. Parma: European Food Safety Authority; 2014(http://www.civ-viande.org/wp-content/uploads/2015/03/743e.pdf, accessed 21 March 2018).
12. Moran-Gilad J. Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS. Global Microbial Identifier initiative's Working Group 4 (GMI-WG4).Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infect Dis, 2015;15:174.
13. Endrullat C, Glökler J, Franke P, Frohme M.Standardization and quality management in next-generation sequencing.ApplTranslGenom. 2016;10: p. 2-9.
14. Franz E, Gras LM, Dallman T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Current Opinion in Food Science. 2016;8: 74-79.
15. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, Harmsen D, Hendriksen RS,et al. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response.Emerg Infect Dis, 2012. 18(11): e1.
16. PallenMJ. Microbial bioinformatics 2020.MicrobBiotechnol. 2016;9(5): 681-6.

3. Implementing WGS as a tool for public health in low- and middle- income countries: the main challenges

There are several key challenges in implementing WGS as a tool to support FBD surveillance and outbreak response and using the information generated in public health action in low- and middle-income countries.

3.1 Infrastructure

WGS should be implemented for public health purposes only where there is already a basic epidemiology, surveillance and food monitoring and testing infrastructure in place. Many low-income countries are still developing basic surveillance and food monitoring systems. A second prerequisite is the presence of functional authorities or agencies that can act on the data produced through WGS. Where there is no such infrastructure, it will be essential to establish an effective food control system that includes routine collection and analysis of clinical, food and environmental samples.

One of the main challenges is ensuring sufficient laboratory capacity to support the surveillance of FBDs in humans. The collection and sequencing of human specimens needs to be timely and reliable. If the food monitoring system is generating only a few isolates, it will be even more important to have strong epidemiological capacity to identify the source of a FBD outbreak. Cases will need to be identifiable either through the surveillance system or through active case-finding, and would need to be interviewed to obtain a history of food consumption. These food histories should then be compared to generate a possible hypothesis about the source of the illness, which is then tested in an analytical epidemiological study. This can help guide investigators in food trace-back activities and sampling of food from the marketplace. Training and retaining epidemiologists to do this is a challenge in most developing countries.

An additional challenge is meeting shipping requirements for reagents. Many sequencing reagents have a limited shelf-life and must be frozen or refrigerated; some can be stored at ambient temperatures. The availability and cost of appropriate cool- or cold-chain shipping methods and storage may significantly affect the ability to maintain the quality of reagents. Ensuring that an adequate supply of all commodities is available at all times is a continual challenge for laboratory supply chain management; if any one of the components of a technique is not available, the entire test cannot be performed *(1)*. Laboratories in developing countries must rely on existing supply pipelines or create new ones, such as interregional warehouses, that could solve part of their problems. Some modern sequencing platforms are now minimizing the commodities needed (e.g. MiniON); this approach shows promise for simplifying supply chain requirements for developing countries. Futhermore, continuity of power supply, proper environmental conditions inside laboratories, and regular maintenance of equipment must be guaranteed.

3.2 Costs

3.2.1 Overall cost

Cost is one of the most important considerations for developing countries. Even if the cost has been diminishing over the years, WGS is still very expensive compared with other current techniques, such as PFGE. At the beginning of WGS implementation when it is used as a complementary technique, adding this cost to routine surveillance may be impossible in some countries. Eventually, as WGS replaces existing techniques (from biochemical identification to serotyping and PFGE for subtyping), the cost will be more affordable, and may even be cheaper than current techniques. In most developing countries, the priority is still to strengthen the surveillance system rather than invest in new technology *(2)*.

The additional costs associated with WGS will depend on the existing level of complexity in the surveillance and response systems. Where effective systems are in place, WGS is likely to reduce the total costs, since it can replace several traditional typing methods. For countries with limited laboratory surveillance in place, implementing WGS will incur significant additional costs. However, many of these costs would be incurred with the implementation of conventional techniques as well. Therefore, costs are relative and difficult to compare between countries.

A full economic cost-benefit analysis of applying WGS routinely in the food safety and health sectors is not yet available; any such analysis is likely to be country- or region-specific. Available cost estimates often focus on the financial benefits of replacing multiple tests with a single sequencing assay *(3, 4)*. Current cost estimates for establishing sequencing capabilities for public health laboratories transitioning from PFGE to WGS vary between US$100 000 and US$700 000, depending on the throughput of isolates and the need for draft or complete genomes. Without donor support, this is likely to be prohibitive for many countries. Estimates of the cost of consumables (e.g. reagents and computational requirements) are similarly varied, and depend on the needs of the specific laboratory *(5)*. Unfortunately, these costs are often higher in developing countries because of the higher costs of shipping, customs formalities, and profit margins for local companies and distributors *(2)*.

3.2.2 Consumables

One of the challenges of implementing any laboratory technology in developing countries is procuring equipment and reagents. With traditional methods, laboratories will need to buy and maintain stocks of organism- or method-specific reagents, such as diagnostic antisera, antimicrobial susceptibility testing discs, and selective and non-selective media. WGS requires less complex and heterogeneous reagents, allowing the overall management of laboratory commodities to be simplified.

While the cost of sequencing has declined over time with the development of less expensive technology and platforms, the equipment and reagents are still expensive for many countries, and may be a barrier to implementation for low-income countries. The cost of sequencing is also a sensitive issue where researchers and scientists rely on small amounts of funding to maintain critical public health and food safety work *(6, 7)*.

This may contribute to a cycle of disadvantage: many peer reviewers may expect or only accept the results of WGS and opportunities to publish may therefore be restricted *(8)*.

Many next-generation sequencing platforms rely on reagents sold as kits. These provide a convenient and easy to use format, but are often expensive. Many laboratories in developing countries have a history of creativity and innovation in adapting sequencing methods to their resources, such as substituting equipment, re-using disposable materials, and producing homemade kits *(8)*. When kits contain proprietary or patented materials, however, it may not be possible to find low-cost alternatives. When selecting laboratory commodities, developing countries typically focus on whether or not the instrument is part of a closed or open system (closed systems require specific brands of reagents). Closed systems create dependence on a single source, but also typically ensure high reagent quality *(1)*. Dependence on a single supplier for their proprietary products has negative economic consequences, especially in resource-limited settings where the recurring costs of running and maintaining equipment can be burdensome. One possible solution is for regional centres to work together to negotiate bulk pricing to lower the overall cost. However, this strategy may not work in every situation and is not practical for an isolated laboratory.

3.2.3 Personnel

The cost of sequencers and supplies might seem prohibitive to many developing countries, but needs to be considered in comparison with the cost of training and maintaining expertise for traditional methods. WGS is a method that is simple to learn and, with automated analytical tools, can be performed with limited staff. In contrast, with traditional methods, it is not feasible for staff members to achieve and maintain sufficient expertise to deal with all or even most foodborne pathogens.

A key component of the transition to WGS in low-income countries will be training. There are multiple online resources regarding WGS data generation and analysis, many of which have been largely driven by academic research organizations. Several academic organizations run residential and non-residential courses in WGS and, while many of these are tailored to the tools used by specific institutions, the skills, methods and resources are highly transferable. Some examples of international training courses are the Genome Campus Advanced Courses, supported by the Wellcome Trust, week-long courses by the US Centers for Disease Control and Prevention (US CDC) on the use of whole genome MLST and PulseNet, and a combined effort from the University of Maryland's Joint Institute for Food Safety and Applied Nutrition (JIFSAN) and the US Food and Drugs Administration (FDA) Center for Food Safety and Applied Nutrition (CFSAN). These courses are accessed through a competitive application process and provide training and support with international experts.

3.3 Bioinformatics

Bioinformatics analysis is critical for the use of WGS data in surveillance. Many countries may benefit from the use of open source or free bioinformatics tools (see Section 2); however, use of these resources requires fast and stable internet connections and the ability to transfer genomic data. Such connections are not always available, particularly in Africa *(14)*.

The output from WGS needs to be in a format that is useful for food safety and public health officials. While dendrograms, SNP matrices and minimum spanning trees can be useful in outbreak investigations, they are not useful for long-term surveillance of foodborne pathogens. Ideally, the outputs from WGS should allow rapid comparison of the sequence data with epidemiological data gathered from the routine surveillance system or from food history interviews. It is only by bringing the laboratory and epidemiological data together quickly that outbreaks and potential sources can be identified. Strain nomenclature, which is crucial for coupling of surveillance networks, is a construct devised to classify an isolate as a whole, that is to place it in some designated category within the species. Currently, the best option for such a nomenclature is allele-based typing, as explained in Section 4, although universally agreed allele typing schemes are not yet available for all the different pathogens. However, SNP-based typing can always be used in parallel to derive high-resolution comparisons (for example, in specific outbreak situations or for regulatory decision-making).

The use of established core-genome MLST schemes allows a common nomenclature to be introduced for genetically related strains *(15-17)*. It is not yet clear in how many alleles two genomes may differ to call them (close to being) identical. The same problem applies when comparing two genomes using SNP typing *(18)*. However, identifying the genetic distance allows unbiased comparisons for different core-genome or whole genome MLST schemes, as well as the definition of thresholds, by studying collections of epidemiologically linked and non-epidemiologically linked strains *(19)*. More established typing schemes for pathogens and cut-off values for typing schemes have to be established, leading to international reference databases with genetic and metadata.

3.4 Data sharing

3.4.1 Harmonization

The ultimate goal of data sharing is a globally accessible online repository and bioinformatics tool that can automatically analyse, match and interpret WGS data. In addition to the technical challenges, there are legal and ethical hurdles to be overcome.

One of the main prerequisites for useful data sharing is harmonization both of the methods used in data production and of the data and associated metadata reported to the sharing platforms. In an era when data generation is becoming more affordable and the number of WGS studies around the world is growing, this is essential. Compliance of shared data and metadata with relevant standards and having adequate reporting systems in place will ensure that:

- data providers can be guided through the reporting processes;
- data and associated metadata are consistently and adequately described;
- data and associated metadata are validated;
- data become discoverable;

- data are more reproducible;
- data are interoperable and usable.

As a tool for the detection, investigation and control of international outbreaks, WGS is unsurpassed with proper standardization. If the raw sequences of all pathogens tested in one country from any source are shared in the public domain, they can provide useful information to public health scientists investigating outbreaks in different countries. Sequencing information may also be shared confidentially through international surveillance networks such as PulseNet International. In this way, public health needs may be served while at the same time the confidentiality of the countries submitting the data is protected.

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing public initiative for sharing molecular data for research. Provision of nucleotide sequence data to INSDC has become a central step in disseminating research findings to the scientific community. INSDC has a uniform policy of free and unrestricted access to all the records in their databases and the INSDC partner databases – the DNA databank of Japan, the European Nucleotide Archive of the European Molecular Biology Laboratory, and the US GenBank and SRA systems – work with publishers of scientific literature and funding bodies to ensure compliance with this principle and to provide submission, storage and data access tools that go hand-in-hand with the publications. INSDC covers a broad spectrum of data from raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations. INSDC serves both as a forum for rapid sharing of sequence data and as the database of record, making it appropriate for both rapid and urgent surveillance and outbreak investigation applications, such as typing, AMR prediction and epidemiological analysis, and critical for the longer-term upkeep of reference data resources.

3.4.2 Data ownership

The perception that data produced in low-income countries are used by high-income countries without due credit triggers ethical concerns. Some scientists may choose to hold onto their data for fear that others may use them unethically. A functioning and effective global WGS data-sharing mechanism will only be possible if all users can be sure that sharing their data will not work against them. The legal ownership of WGS data produced from isolates collected by different institutes (hospitals, food authorities, public health agencies) and from different sources (patients, retailers, companies, etc.) is an issue. This may be a particular concern for data on isolates from commercial parties (collected for example by food safety authorities). There are a few global legal frameworks, such as the International Health Regulations (2005) (IHR 2005) and the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. Since the data contain metadata, privacy and confidentiality remain key issues. The IHR 2005 provide a structure for sharing metadata that contain information on individuals, which is critical in managing serious outbreaks, showing that it is not entirely impossible to handle such sensitive information at the global level. However, the IHR 2005 are purpose-specific and not easily applicable for other purposes. Developing a globally harmonized legal framework for all types of WGS data sharing for all countries may not be realistic given the complexity of each country's

legal structure and context. At the same time, the amount of WGS data generated is so large that it is no longer realistic to share data on a trust basis. Therefore, in order eventually to implement global sharing of WGS data (which has widely acknowledged benefits), it will be necessary for international organizations and their member countries to reach a common understanding.

While this document focuses on WGS of pathogens for surveillance and outbreak response, there is also a need for consideration of privacy and ethics in relation to all human data, including demographic and phenotype data associated with disease surveillance. All published patient data should contain no identifiers that can be traced to individuals; this requires compliance on the part of the research team and needs to be facilitated technically in the data capture, management and integration systems. Some countries have legislation that prohibits the export of human data, in particular, and in some cases other biological samples and data. These concerns are coming increasingly to the fore in discussions around the use of commercial clouds for data storage and processing. This supports the argument for federated systems for data storage, so that sharing of data "across borders" requires specific access or processing tools and technical skills. Efforts to manage this process for human genetic data have progressed under the umbrella of the Global Alliance for Genomics and Health (http://genomicsandhealth.org). Individual datasets will also need to be managed appropriately to ensure data integrity, reliable storage and backup, and efficient access mechanisms (reliable Internet access and speeds).

3.4.3 Metadata and ontology

Metadata are crucial for data sharing, as is data harmonization. Data need to be well curated, and FAIR (findable, accessible, interoperable and reusable), as well as directly comparable across sites. There are several community efforts to use ontologies and PhenX measures *(20)* to ensure that clinical and genomic data are harmonized, but uptake is not yet global and there are still several gaps in ontologies for certain pathogens or phenotypes. For example, there are different ontologies to describe the metadata for different disciplines, such as the gene ontology (GO), describing gene function and cellular location, the Standardized Nomenclature of Medicine, for human medicine, and the Standardized Nomenclature for Veterinary Diagnoses and Operations. Without a detailed ontology, there is a risk that different parties will record the same information in slightly different ways, thereby limiting analyses. Standardized communication protocols will be needed to allow real-time exchange and monitoring. This in itself can be a huge task, distant from, but essential for effective and coordinated surveillance efforts. This challenge is compounded by the dynamic nature of microbes, whereby common descriptors used to categorize pathogens – species, subspecies, strain and serotype – fail to represent accurately the continuum of variation seen when whole populations are sequenced. New lexicons are needed to describe or categorize microbial diversity to suit the questions being addressed.

3.4.4 Data analysis

Data harmonization, submission to public repositories and federated sharing all require processes and systems to be in place that allow seamless data sharing. These include data storage and analysis platforms and application program interfaces (APIs) for querying, accessing or processing the data. Several pathogen

outbreak response tools have been developed and are discussed below. However, it is unclear which system would be best suited for surveillance systems looking to put genomic epidemiology into practice, or how to go about setting up the infrastructure for the system. In some cases, sequencing and data analysis are conducted at separate sites, making data transfer an important component of the data management environment. Analysis of surveillance and pathogen outbreak response data, including rapid identification of drug resistance, requires appropriate analysis and visualization tools, as well as expertise in using them and interpreting the results. As mentioned previously, technical skills are also required to provide support for data access or submission to public repositories. These skills are often lacking in low- and middle-income countries.

In general, efforts are needed to improve the skills of researchers in low- and middle-income countries in data analysis, in order to make them internationally competitive and to facilitate filing and exploitation of intellectual property. The surveillance and sequence data have excellent potential for commercial application, which may mean that some researchers are reluctant to share. However, if researchers were equipped to analyse the data rapidly and file relevant patents, the data could be shared sooner. Patents can then be licensed to companies with the ability to respond rapidly in developing new diagnostics or therapeutics. Here again, however, commercialization potential favours experienced, well-resourced laboratories.

3.4.5 Trade implications

There are a number of additional negative implications of data sharing in the commercial sector, particularly for the trade of food and animals. Recent outbreaks have highlighted the advantages of early detection of disease outbreaks and of regular monitoring of the health status of national livestock populations. The World Organisation for Animal Health (OIE) (http://www.oie.int/) has developed guidelines for countries that trade in animals and animal products on implementation of animal disease surveillance systems *(21)*, and the Food and Agriculture Organization of the United Nations (FAO) (http://www.fao.org/home/en/) has developed the Emergency Prevention System for transboundary animal and plant diseases. Implementation of such systems and regular monitoring of livestock allow early detection of potential threats, thus benefiting the health of the livestock and permitting a rapid response to minimize economic impact. However, while these increase confidence in the food industry, the public availability of this information may foster fear and distrust of trade from the affected country or loss of confidence in the company producing the affected product. This can have serious long-term economic consequences for the country or company, thus dissuading governments and companies from releasing sensitive information through data sharing. If an outbreak is traced back to an exported commodity or animals from a particular country or company, this could negatively impact their economies or profits respectively, and, of course, result in the devastation of their livestock. While rapid availability of outbreak data can reduce the economic loss to some extent, this creates a predicament for the food provider in rapidly releasing information that will benefit health, but that may be harmful to their industry.

3.5 References

1. USAID Deliver project, Task ORder1. Guidelines for managing the laboratory supply chain: Version 2. Arlington, USAID Deliver project, Task Order1. 2008. (http://pdf.usaid.gov/pdf_docs/Pnadg882.pdf, accessed 21 March 2018)
2. Helmy M, AwadM, Mosa KA. Limited resources of genome sequencing in developing countries: challenges and solutions. Applied and Translational Genomics. 2016(9):15-9.
3. Expert opinion on whole genome sequencing for public health surveillance.Stockholm: European Centre for Disease Prevention and Control; 2016.
4. ECDC roadmap for Integration of molecular typing into European-level surveillance and epidemic preparedness. Version 2.1, 2016-2019. Stockholm: European Centre for Disease Prevention and Control; 2016.
5. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11(5):759-69.
6. Applications of whole genome sequencing in food safety management. Rome: Food and Agriculture Organization of the United Nations: 2016.
7. Technical meeting on the impact of whole genome sequencing (WGS) on food safety management: within a One Health approach. The 9thmeeting of the Global Microbial Identifier (GMI9) 23–25 May 2016 Rome, Italy. Rome; Food and Agriculture Organization of the United Nations; 2016 (http://www.globalmicrobialidentifier.org/news-and-events/nyheder/nyhed?id=2119C9F8-B62A-4575-B326-DFC96B519480, accessed 21 March 2018).
8. Coloma J, Harris E. Innovative low cost technologies for biomedical research and diagnosis in developing countries. BMJ. 2004;329(7475):1160-2.
9. Pallen MJ. Microbial bioinformatics 2020. Microb Biotechnol. 2016;9(5):681-6.
10. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E et al. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infect Dis. 2015;15:174.
11. Endrullat C, Glokler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. ApplTranslGenom. 2016;10:2-9.
12. Franz E, Gras LM, Dallman T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Current Opinion in Food Science. 2016;8:74-9.
13. 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.
14. Karikari TK, Quansah E, Mohamed WM. Developing expertise in bioinformatics for biomedical research in Africa. ApplTranslGenom. 2015;6:31-4.
15. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W et al. Core genome multilocus sequence typing scheme for high-resolution typing of Enterococcus faecium. J ClinMicrobiol. 2015;53(12):3788-97.
16. Kohl TA, Diel R, Harmsen D, Rothganger J, Walter KM, Merker M et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J ClinMicrobiol. 2014;52(7):2479-86.
17. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F et al. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Listeria monocytogenes. J Clin Microbiol. 2015;53(9):2869-76.
18. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728-36.
19. Kluytmans-van den Bergh MF, Rossen JW, Bruijning-Verhagen PC, Bonten MJ, Friedrich AW, Vandenbroucke-Grauls CM et al. Whole-genome multilocus sequence typing of extended-spectrum-beta-lactamase-producing Enterobacteriaceae. J ClinMicrobiol. 2016;54(12):2919-27.
20. Hendershot T, Pan H, Haines J, Harlan WR, Marazita ML, McCarty CA et al. Using the PhenX Toolkit to Add Standard Measures to a Study. CurrProtoc Hum Genet. 2015;86(1.21):1-17.
21. Kloeze H, Mukhi S, Kitching P, Lees VW, Alexandersen S. Effective animal health disease surveillance using a network-enabled approach. TransboundEmerg Dis. 2010;57(6):414-9.

4. The current state of WGS technology and the supporting bioinformatic tools

WGS and the bioinformatic methods for analysing the data produced are evolving rapidly. This section considers the current state of the WGS technology being used in public health, including the instrumentation and the bioinformatics.

4.1 WGS instrumentation and capacity

The wide range of sequencing instruments available provide many options for FBD disease surveillance, each with its own strengths and weaknesses. Technical characteristics and questions regarding infrastructure as well as costs and availability of instruments and reagents are important factors that need to be considered in tandem, especially in the developing world. More complicated – but no less important – is the question of how sequencing should be implemented, whether it should be carried out centrally or at a more local level, and what these choices mean for the selection of instrumentation.

One of the first things that must be considered is which technology is most applicable to the task. Important characteristics, such as throughput, run time, instrumentation and continuing service costs, reagent costs, and infrastructure and staffing needs, must be taken into account. For example, the monitoring of some endemic diseases, such as malaria, may require a regional sequencing infrastructure capable of handling hundreds of unique samples each week. On the other hand, sporadic events, such as hospital outbreaks of *Salmonella,* may require more local centres capable of examining fewer samples more quickly. In cases of highly virulent or rare outbreaks, portable sequencing equipment that can be used on-site may be most appropriate. The following sections will discuss the various technical and cost parameters of common sequencing platforms, in order to assist in this decision-making process. It should be kept in mind that sequencing technology is evolving rapidly and new, more accessible platforms could well emerge in the near future.

4.1.1 Short-read platforms

The majority of WGS platforms fall under the category of short-read sequencers. These sequencers have a maximum read length of 1000 bases, but a more typical base range is between 50 and 400 *(1)*. The error rate for these platforms is quite low, with accurate nucleotide calling in excess of 99%. Traditionally, short-read platforms are further broken down into sequencing-by-synthesis (SBS) and sequencing-by-ligation (SBL) methods. There are currently two available SBL platforms; however, these instruments account for only a small percentage of sequencers in use because SBS has a higher output and is cheaper. SBL will, therefore, not be discussed further.

There are two relevant SBS platforms in use: the Illumina suite of instruments and, to a lesser extent, the Ion Torrent suite.

The Illumina suite of instruments relies on incorporating fluorescently-labelled nucleotides in an elongating DNA strand. The nucleotides are modified such that only one base is incorporated at a time. As each base is incorporated into the elongating strand, the instrument identifies the nucleotide base in either two (MiniSeq, NextSeq) or four detection channels (MiSeq, HiSeq). This approach gives the instrument very high accuracy and throughput. There are three broad types of illuminainsutruments.

- *HiSeq instrument*. These machines offer the lowest cost per gigabase (Gb) of any currently available platform. However, while the throughput of these instruments is quite useful for larger genomes (such as the three billion nucleotide human genome), their application to pathogen sequencing of a few million bases is not practical. For example, one lane of the HiSeq 2500 generates approximately 60Gb of paired-end sequencing data of 125 bases in length or approximately 1 human genome at 18X coverage. This is equivalent to more than 66 *E. coli* genomes at similar coverage, far beyond what can reasonably be expected (or needed) at a typical pathogen sequencing centre. This instrument's capabilities are more appropriate for activities carried out at human research institutions and eukaryotic genome centres.
- *NextSeq instrument*. Unlike the HiSeq instrument, this device does not require a minimum number of samples, making it a more practical option for sequencing centres that do not expect a regular influx of samples. The NextSeq provides a per lane throughout similar to that of the HiSeq 2500, and at similar costs per Gb sequenced. This makes the instrument a good fit for an academic sequencing core or a regional centre focused on human sequencing. However, like the HiSeq, its throughput is likely to be beyond what is needed for a beginning pathogen sequencing centre.
- *MiSeq and MiniSeq instruments*. These have 0.5 to 15 Gb and 1.5 to 7.5 Gb output, respectively. Like the NextSeq, neither of these instruments requires a minimum number of samples. From a technical perspective, these instruments are therefore well suited for pathogen sequencing. The lowest throughput setting (single reads of 36 bases in length on the MiSeq) can generate one *E. coli* genome at 100X coverage in as little as 4 hours. With one of these instruments, a regional sequencing centre could comfortably sequence 96 bacterial genomes per week. While the cost per Gb of these instrument is higher than that of the higher throughput platforms, the initial instrument cost is substantially lower.

Less widely used, but in a similar niche as the MiSeq and MiniSeq, is the suite of instruments from Ion Torrent (ThermoFisher). Unlike the Illumina products, the Ion Torrent instruments detect bases through the release of a hydrogen ion during strand elongation rather than an optical signal. This method provides one notable advantage over the Illumina suite: the Ion Torrent suite offers the fastest sequencing time (as little as 2.5 hours) of any currently available instrument. However, this method of detection also leads to a higher error rate than the Illumina suite, especially in homopolymeric regions, leading to more difficult *de novo* assemblies.

While the initial cost of the suite of instruments from Ion Torrent is comparable to that of the Illumina suite, the overall cost per billion bases sequenced is somewhat higher, owing to the lower number of

reads produced per run. Furthermore, there is a smaller community of active users and publicly available software developers for the Ion Torrent and its downstream data analysis, which means fewer options for data pipelines and fewer resources available for troubleshooting. For these reasons, the Ion Torrent suite of products is best reserved for targeted sequencing projects (e.g. 16s RNA, virulence marker identification and transcriptome profiling).

4.1.2 Long-read platforms

Long-read sequencing platforms are typically considered to be capable of generating average read length in excess of 10 kilobases (kb), the generally accepted minimum read length for high-quality long-read assemblies. There are two currently available long-read platforms: the Pacific Biosciences (PacBio) suite and the Oxford Nanopore Technologies (ONT) suite. Both of these platforms can generate both short and long reads, the final read length being defined by the input DNA fragments rather than the instrument itself.

The PacBio suite consists of the RS II and Sequel instruments. Like short-read platforms, PacBio relies on a sequencing-by-synthesis approach, in which fluorescently-labelled nucleotides are detected as they are incorporated into an elongating DNA strand. The PacBio approach is called single-molecule real-time (SMRT) sequencing, owing to the fact that single DNA molecules are monitored with no pausing steps for interrogation. SMRT technology allows reads in excess of 60kb, but with error rates as high as 15%. As a result these long-read *de novo* assembly strategies must achieve higher cumulative coverage (approximately 120X) to overcome this error rate. In terms of yield, the RS II can generate approximately 1 Gb of data and the Sequel approximately 5 Gb per SMRT cell; however, only about half that yield contains reads over 10 kb. This means that two, long-read *E. coli* genomes can be generated at approximately 120X coverage per cell on the RSII and ten on the Sequel, making them well suited to a low throughput sequencing centre.

While the output from short-read platforms can result in genome assemblies that are nearly complete, some gaps are still expected because of the shorter length of the sequencing reads used. The long-read length of PacBio offers an important and distinct advantage of SMRT sequencing, in that the reads are able to produce a high-quality genomic sequence that typically captures all of the genetic material seen (e.g. closed genome); this is important when regulatory authorities need to have a genomic sequence that is as complete as possible. The main drawbacks of the PacBio suite are the cost of the instrument (more than US$ 350 000 for the Sequel), the cost of reagents, and the infrastructure required. Both RS II and Sequel have footprints many times larger than most other sequencers and require a continuous supply of nitrogen, potentially limiting their application in developing countries.

The ONT suite of instruments, including the MinION and PromethION, use a unique sequencing method: single DNA molecules translocate through a biological pore, and small perturbations of current passing through the pore can be interpreted in terms of bases. The MinION instrument has a 3 cm x 10 cm footprint and is more portable than other sequencing platforms; and without the need for significant fluids or optics, the instrument cost is negligible. The throughput per MinION instrument is between 5 and 10 Gb per 48-hour run (with the option of shorter runtimes), while the PromethION is expected to generate as much

as 60Gb per flowcell. While this throughput may be somewhat high for a pathogen sequencing centre, several benefits such as iterative loading, shorter runtimes, and real-time base calling abrogate the need for significant multiplexing. The limitations of the ONT suite include a biased error profile, with indels being especially problematic in repetitive regions. Reagent costs are also higher than the other platforms discussed here, though the low instrument cost and upcoming "flongle" flowcells may overcome this.

4.1.3 Summary

In selecting an appropriate sequencing technology, the following factors will need to be taken into account.

(1) The cost of both the sequencing machine and consumables required for the sequencing platform to be used.

(2) The running costs of the machine in relation to the number of samples expected to be sequenced each week.

(3) The ease (or availability) of receiving shipments needed to support the continued use of the sequencing platform. It makes little sense to purchase equipment that cannot be easily serviced or maintained.

(4) The need for draft or closed genomes.

(5) Whether new technologies not addressed in this paper would be more appropriate given that the sequencing landscape and its myriad options are evolving rapidly.

A summary of the various sequencing instruments with their technical specifications and costs is given in Table 4.1.

TABLE 4.1

Technical specifications and cost of available sequencers

Platform	Read length	Yield (Gb)	Run time	Instrument cost	Annual contract	Cost per Gb	Disadvantages	Advantages
Illumina MiniSeq	50–150 bp	1.6 to 7.5	7–25 hours	US$ 50 000	US$ 5000	US$ 200–400	High cost per Gb	Low instrument cost, established technology, low error rate
Illumina MiSeq	75–300 bp	0.5 to 15	4–56 hours	US$ 99 000	US$ 14 000	US$ 250–2000	High cost per Gb	Low instrument cost, established technology, low error rate
Illumina NextSeq	75–150 bp	16 to 120	15– 29 hours	US$ 250 000	US$ 32 000	US$ 33– 43	High instrument cost	Low cost per Gb, established technology, low error rate
Illumina HiSeq2500	36–125 bp	9 to 500	7 hours to 11 days	US$ 690 000	US$ 75 000	US$ 30– 230	High instrument cost, need for deep multiplexing	Low cost per Gb, established technology, low error rate

(TABLE 4.1 Continue)

Platform	Read length	Yield (Gb)	Run time	Instrument cost	Annual contract	Cost per Gb	Disadvantages	Advantages
Illumina HiSeq3000/4000	50–150 bp	105 to 750	1–3.5 days	US$ 740 000–900 000	US$ 81 000	US$ 22– 50	High instrument cost, need for deep multiplexing	Low cost per Gb, established technology, low error rate
Illumina HiSeq X	150 bp	800 to 900 per flow cell	< 3 days	US$ 1 000 000	US$ 93 000	US$ 7– 10	High instrument cost, need for deep multiplexing, limited compatibility	Low cost per Gb, established technology, low error rate
Ion PGM	200–400 bp	0.03 to 2	3.7–23 hours	US$ 49 000	US$ 5000–10 000	US$ 400– 2000	Not able to do paired-end sequencing, poor homopolymer performance, high cost per Gb	Rapid sequencing run
Ion Proton	Up to 200 bp	Up to 10	2–4 hours	US$ 224 000	US$ 20 000–30 000	US$ 80	Not able to do paired-end sequencing, poor homopolymer performance	Low cost per Gb, rapid sequencing run
Ion S5	200–400 bp	0.6–8	2.5–6 hours	US$ 65 000	US$ 9000–18 000	US$ 80–500	Not able to do paired-end sequencing, high cost per Gb	Rapid sequencing run
Pacific BioSciences RS II	~20 kb	~1	4 hours	US$ 695 000	US$ 84 000	US$ 1000	13% single pass error rate, very high cost per Gb, high instrument cost	Very long read lengths, can sacrifice length for accuracy, rapid run time
Pacific BioSciencesSequel	~20 kb	~5	4 hours	US$ 350 000	US$ 20 000	US$ 1000	13% single pass error rate, very high cost per Gb, high instrument cost	Very long read lengths, can sacrifice length for accuracy, rapid run time
Oxford Nanopore MK 1 MinION	Up to 200 kb	Up to 10	Up to 48 hours	US$ 1000	0	US$ 100–400	10% single pass error rate, increased indel errors in repeat regions, high cost per Gb	Very low instrument cost, portable

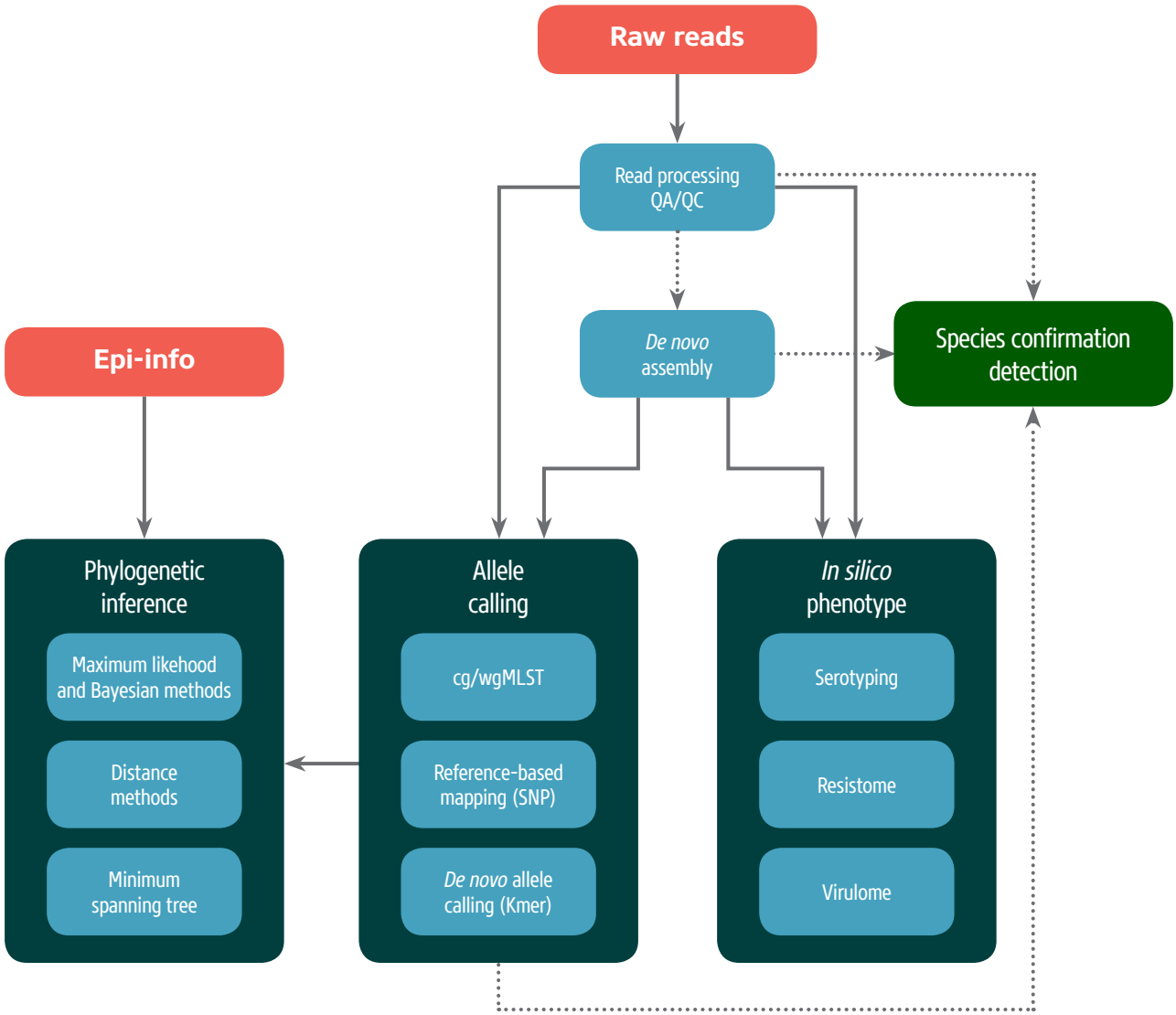
Table adapted from refs 2 and 3

4.2 Bioinformatics of WGS data

Extracting all significant information from several million sequence reads produced by modern sequencers requires considerable computing time and efficiency. In recent years, a number of bioinformatic solutions have been developed to address this challenge. Bioinformatic analyses often involve guiding files through a series of data transformations, called a pipeline or workflow. Routine data analyses can be performed by trained technicians, so-called e-lab technicians. A schematic representation of possible pipelines for using WGS in microbial identification, characterization and typing is shown in Figure 4.1.

FIGURE 4.1

Schematic representation of WGS pipeline



4.2.1 Quality assurance, quality control and read preprocessing

At the beginning of the workflow, quality assurance (QA) and quality control (QC) measures should be implemented to ensure consistent, high-quality comparable genomic data. Software (such as FASTQC *(4)* and SAMStat*(5))* provide a simple way to perform quality control checks on raw sequence data, and provide a flexible set of analyses (e.g. per base sequence quality, nucleotide composition, read-length distribution, base quality distribution) that can be used to estimate the sequence quality and to identify possible errors before proceeding with further analyses. In both cases, the output is a single and easy-to-read Web page, which can be interpreted by either sequencing technicians or bioinformaticists. In addition, operations can be performed automatically to allow a quality check of individual processing steps in large analysis pipelines.

Once the quality of the raw data has been evaluated, reads usually undergo a preprocessing step during which they are "cleaned" to remove low quality data, including any adaptor sequences inserted during library preparation. A few read processing software packages are available; TRIMMOMATIC *(6)* was designed to handle paired-end data. Another software package able to perform automatic QA, QC and read processing is INNUca*(7)*. After the user provides cut-off values for the QA and QC, INNUca will process raw FastQC reads, performing coverage calculations before and after read quality analysis and trimming, *de novo* draft genome assembly and validation, species confirmation and contamination testing.

4.2.2 Species identification

In both surveillance and outbreak investigation of foodborne pathogens, one of the earliest critical steps is the correct identification of the microorganisms of interest at the species level. Analysis of the 16S gene has traditionally been used to determine bacterial species, and a number of annotated databases – such as Greengenes*(8)*, RDP *(9)* and SILVA *(10)* –have catalogued 16S genes from a large number of species. Species information can be extracted rapidly from raw or processed reads, using approaches involving exact alignments of k-mers, such as KRAKEN *(11)*, k-mer distribution, CGE KmerFinder by the Center for Genomic Epidemiology (CGE) *(12)*, or MinHash dimensionality-reduction technique, as implemented in MASH *(13)*. Similarly, speciation can be performed on assembled genomes using either MASH or other fast clustering methodologies such as GScompare *(14)*, which calculate genomic signature distances among oligonucleotide frequencies from different sequences. Finally, data from both WGS typing and MLST can be placed within a phylogenetic tree of known samples, resulting in an implicit confirmation of the species of the samples.

4.2.3 *in silico* typing and phenotype prediction

While several thousands of genomes have been sequenced for some foodborne pathogens, it is still critical to be able to link the sequenced isolates to data in traditional typing databases, in order to place the isolate in an appropriate historical context, improving response to public health events. Traditional molecular typing by MLST still offers a valuable curated way of typing, since there are large databases of MLST genes covering a number of different species *(15, 16)* and MLST can be easily extracted *insilico* from the WGS data. *In silico* MLST analysis of bacterial isolates using WGS data can be performed from draft-genome assemblies using MLST 2.0 *(17)*, CGE Web-based interfaces using MLST 1.8 *(18)*, and directly on PubMLST (which uses BIGSdb platform), Enterobase websites, or by commercial software packages such as CLC Genomic Workbench, Bionumerics and Seqsphere+. Both ReMatCh *(19)* and SRST2 *(20)* use a mapping methodology to perform rapid *in silico* typing that includes not only MLST but also the presence or absence of specific accessory genes. This makes them suitable tools for phenotypic prediction of relevant microbiological features, such as antimicrobial or virulence-associated genes. Some tools and databases available for predicting the antimicrobial resistance and virulence status of certain bacterial species are listed in Table 4.2. These databases can be searched using specific tools, such as the CGE ResFinder *(21)*, the CGE VirulenceFinder *(22)*, Resistance Gene Identifier (RGI) *(23)*, or the BLAST algorithm.

TABLE 4.2

Publicly available antibiotic resistance and virulence databases

Antimicrobial resistance databases	Virulence databases
CARD (https://card.mcmaster.ca/) (*23*)	VFDB (http://www.mgc.ac.cn/VFs/) (*24*)
RAC (http://rac.aihi.mq.edu.au/rac/) (*25*)	PATRIC (https://www.patricbrc.org/) (*26*, *27*)
ResFinder (https://cge.cbs.dtu.dk/services/data.php) (*21*)	Victors (http://www.phidias.us/victors/) (*28*)
ARDB (https://ardb.cbcb.umd.edu/) (*29*)	PHI-BASE (http://www.phi-base.org/) (*30-33*)
NDARO https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/	MvirDB (http://mvirdb.llnl.gov/) (*34*)

WGS data can also be used to predict serotyping information for certain bacterial species by looking for specific genes or by extrapolating from a population structure. WGS-based O and H typing of *Escherichia coli* can be inferred by user-friendly, freely available data analysis Web tools, such as CGE SerotypeFinder*(35)*. Similarly, serotyping of *Salmonella enterica* can be performed through the Web tools SISTR *(36)* and SeqSero*(37)*. In addition to performing serovar prediction by genoserotyping (as performed in SeqSero), SISTR integrates sequence-based typing analyses, such as MLST and core genome MLST, increasing the accuracy of *in silico* serovar prediction. The phylogenetic information extrapolated though core genome MLST or different genome-based methodologies can be applied to predict the serotype of other species such as *Listeria monocytogenes (38)*.

4.2.4 Whole genome molecular typing, allele calling and phylogenetic inference

The groundbreaking advantage of WGS in microbial typing is the possibility to assess variation in hundreds or thousands of targets in the genome simultaneously, rather than focusing on a single or only a few targets, as was the case with MLST, thus providing high-throughput and high-resolution genotyping. Relevant genomic information can be extracted from reads through different allele-calling strategies: *de novo* assembly-based analyses (e.g. gene-by-gene approach), reference-based mapping (e.g. SNP calling) and *de novo* allele calling (e.g. k-mer) *(19, 39)*.

De novo assembly-based analyses

After read processing, several pipelines require the genomes to be assembled into larger continuous sequences or "contigs". Genome assembling is performed on sequence reads using one of the following assembly strategies: overlap, layout, consensus (OLC) and de Bruijn graph *(39)*. Depending on the platform used to produce the reads and the final result (i.e. draft or complete assembly), several software packages can be used. Table 4.3 shows a non-exhaustive review of assemblers.

TABLE 4.3

Some assemblers used in foodborne pathogen WGS pipelines

Assembler	Platforms	Ref
Velvet	Illumina reads	(*40*)
SPAdes	Illumina or IonTorrent reads. It is capable of providing hybrid assemblies using PacBio, Oxford Nanopore and Sanger reads	(*41*)
MIRA	Sanger, 454, Illumina and IonTorrent reads. Can perform hybrid assemblies.	(*42*)
Canu	PacBio and Oxford Nanopore reads. A fork of the Celera Assembler designed for high-noise single-molecule sequencing	(*43*)

Assembled reads can subsequently be analysed using essentially two methodologies: (core) genome alignment and gene-by-gene analysis. Harvest is a software suite including Parsnp, a fast core-genome multi-aligner, and Gingr, a dynamic visual platform for interactive analysis of core-genome alignment, and visualization tools for quickly analysing thousands of intraspecific microbial genomes, including variant calls, recombination detection, and phylogenetic trees *(44)*. In addition to core-genome alignment, genomes can be used for gene-by-gene approaches that revisit the MLST concept but extend it to multiple loci across the whole genome, or use only loci present in a core genome shared by most strains of a given species *(45)*. Species-specific databases and whole genome or core-genome MLST schemas are available in Enterobase (*E. coli*, *Salmonella enterica* and *Yersinia* spp. *(15)*), pubMLST (*Campylobacterjejuni/coli (16)*) and PasteurMLST (*Listeria monocytogenes (46)*). Enterobase uses its specific allele-calling engine after raw reads, submitted by the user through a Web secure access, are passed through defined QA/QC evaluation. In contrast, both pubMLST and PasteurMLST use BISGdb as basic platform, which includes a specific allele-calling methodology, after submission of assembled genomes by a curator.

Several other open-source allele-calling algorithms have been developed for whole genome and core-genome MLST. Genome Profiler allows ad hoc whole genome MLST analysis of a set of bacterial genomes specifically to account for gene paralogy*(44)*. The software chewBBACA is a comprehensive pipeline for creating and validating whole genome and core-genome MLST schemas, providing an allele-calling algorithm based on BLAST score ratio *(47)* that can be run in multiprocessor settings *(48)*. In addition, whole genome and core-genome MLST analyses are implemented in commercial software, particularly Bionumerics (Applied Maths) and RidomSeqSphere+ (Ridom GmbH). Both platforms offer ready validated whole genome MLST schemas for several bacterial species as well as the possibility to develop specific customized schemes. For example, US CDC have created a whole genome MLST scheme for *Listeria monocytogenes* inside the software already implemented within PulseNet (BioNumerics 7.5, Applied Maths), allowing federal, state and local public health laboratories to identify highly related isolates *(49)*.

Reference-based mapping (SNPs)

In reference-based approaches, processed reads are mapped to a reference (a high-quality finished genome) as a basis for the discovery of SNPs that can be used to infer phylogenetic relationships *(39)*. The SNVPhyl (single nucleotide variant phylogenomics), implemented in the Integrated Rapid Infectious Disease Analysis Project (IRIDA; www.irida.ca) platform, is a pipeline for identifying SNPs within a collection of microbial genomes and constructing a phylogenetic tree *(50)*. Rapid bacterial SNV calling and core genome alignments can be performed quite quickly from raw reads and a reference genome using the software Snippy *(51)*, which has been used in Nullarbor, a pipeline dedicated to generating complete public health microbiology reports from sequenced isolates *(52)*. The CGE makes available its Web-based tool, CSI Phylogeny 1.4, which generates an SNP tree using reference-based mapping *(22)*. Reference-based SNP-calling and phylogeny are also implemented in commercial software, such as Bionumerics (Applied Maths) and CLC genomic workbench (Qiagen).

The FDA's Center for Food Safety and Applied Nutrition (CFSAN) provides GenomeTrakr, the first distributed network of laboratories to collect, sequence and share the genome of over 150 000 foodborne pathogens collected from clinical, environmental and food isolates. The database is housed at the National Center for Biotechnology Information (NCBI) and uses NCBI's pathogen detection pipeline *(53)* to produce daily phylogenetic trees that various public health authorities can use to respond to foodborne outbreaks. Furthermore, the CFSAN SNP Pipeline *(54)* is a peer-reviewed method for identifying variants both in outbreak detection and for preventive control analysis *(55)*.

De novo allele-calling

De novo allele callers are developed to deduce phylogenetic relationship among isolates, particularly from raw, unassembled reads, based on k-mers. This methodology is implemented for example in kSNP*(56)*, and is a fast way of finding isolates that are similar to each other. Among the benefits of this approach is that it can find a more ideal reference sequence for use in a reference-mapping approach or show how similar (or dissimilar) a group of isolates are to each other.

4.2.5 Examples of bioinformatic tools

Table 4.4 summarizes some of the bioinformatics tools that are currently being used in public health settings, including for FBD surveillance and outbreak response.

TABLE 4.4

Bioinformatics tools for foodborne pathogens

Category	Tool	Description	Interface	Availability	URL
Database	NCBI Pathogen Detection	Genomic sequences of bacterial pathogens from food, environment and patients. Pathogen clustering and identification for tracking food contamination and helping in foodborne outbreak investigation	GUI	Free access	ncbi.nlm.nih.gov/pathogens/
	Enterobase	An online resource for analysing and visualizing genomic variation within enteric bacteria	GUI	Free access	enterobase.warwick.ac.uk/
	pubMLST	An online resource for analysing and visualizing genomic variation within enteric bacteria	GUI	Free access	pubmed.com/
Multipurpose platform	CGE Toolbox	A suite of Web-based tools and services for pathogen typing and phylogeny construction	GUI	Free access	https://cge.cbs.dtu.dk//services/all.php
	IRIDA	A Web platform to support real-time infectious disease outbreak investigation using genomic data	GUI	Open source	irida.ca/
	BioNumerics	A software platform for biological data management and analysis, including WGS data	GUI	Proprietary	applied-maths.com/bionumerics
	CLC Genomics Workbench	A software package for analysing and visualizing NGS data	GUI	Proprietary	qiagenbioinformatics.com/products/clc-genomics-workbench/
	Geneious	A suite of software tools for molecular biology and NGS data analysis	GUI	Proprietary	basespace.illumina.com/home/
	SeqSphere+	A software package for analysing NGS and Sanger sequencing data to support outbreak investigation and surveillance	GUI	Proprietary	ridom.com/seqsphere/
Analytical pipeline	CFSAN SNP Pipeline	An SNP pipeline developed by Center for Food Safety and Applied Nutrition, US Food and Drug Administration	CLI	Open source	snp-pipeline.readthedocs.io/en/latest/
	LYVE-SET	An SNP pipeline developed by Enteric Diseases Laboratory Branch, US Centers for Disease Control and Prevention	CLI	Open source	github.com/lskatz/lyve-SET
	Snippy	A pipeline for rapid identification of haploid variants and construction of phylogeny using core genome SNPs	CLI	Open source	github.com/tseemann/snippy
	Harvest	A suite of core-genome alignment and visualization tools for quick and high-throughput analysis of intraspecifc microbial genomes	CLI	Open source	harvest.readthedocs.io/en/latest/
	BigsDB	A software tool to store and analyse sequence data for bacterial isolates by extending the principle of MLST to genomic data	CLI	Open source	bigsdb.readthedocs.io/en/latest/
	Nullabor	A pipeline for generating public health microbiology reports from sequenced isolates including sequencing specifics, species identity, subtypes, etc.	CLI	Open source	github.com/tseemann/nullarbor
	Phyloviz	A software tool for analysing and visualizing sequence-based typing and associated epidemiological data	CLI	Open source	phyloviz.net/
Speciality tools	SeqSero	A Web and command line-accessible pipeline for *Salmonella* serotype prediction from raw reads and genome assemblies	GUI, CLI	Open source	denglab.info/SeqSero
	SISTR	A Web-accessible tool for *Salmonella* typing using draft genome assemblies	GUI, CLI	Open source	lfz.corefacility.ca/sistr-app/
	Microreact	A Web-based tool for genomic epidemiology data visualization and sharing	GUI	Free access	microreact.org/

CLI: command line interface, GUI: graphical user interface, NGS: next generation sequencing

4.3 References

1. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017;243:16-24.
2. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11(5):759-69.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.
4. Babraham Bioinformatics. FastQC: A quality control tool for high throughout sequence data [software application]. Cambridge: The Babraham Institute; 2017. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed 10 April 2018).
5. Lassmann T, Hayashizaki Y, Daub CO. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics. 2011;27(1):130-1.
6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
7. INNUENDOCON/INNUca. [website]. San Francisco: GitHub Inc.; 2016. (https://github.com/INNUENDOCON/INNUca, accessed 10 April 2018).
8. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72.
9. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42(Database issue):D633-42.
10. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6.
11. Wood DE Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46. .
12. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H et al. Benchmarking of methods for genomic taxonomy. J Clin Microbiol. 2014;52(5):1529-39.
13. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S et al. MASH: fast genome and metagenome distance estimation using MinHash. J Clin Microbiol. 2014;52(5):1529–39.
14. GScompare:comparing oligonucleotide-based genomic signatures among sequences [website]. Vizcaya: University of the Basque Country (UPV/EHU).(http://gscompare.ehu.eus/, accessed 10 April 2018).
15. Enterobase [website]. Coventry: University of Warwick; 2018. (https://enterobase.warwick.ac.uk/, accessed 10 April 2018)
16. Campylobacter Multi Locus Sequence Typing website [website] In: PubMLST. Oxford: University of Oxford; 2018. (http://pubmlst.org/campylobacter/, accessed 10 April 2018).
17. Scan contig files against PubMLST typing schemes [website]. San Francisco: Github Inc. (https://github.com/tseemann/mlst., accessed 10 April 2018).
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. Machado MP, Ribeiro-Gonçalves B, Silva M, Ramirez M, Carriço JA. Epidemiological surveillance and typing methods to track antibiotic resistant strains using high throughput sequencing. Methods Mol Biol. 2017;1520:331–55.
20. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6(11):90.
21. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640-4.
22. Center for Genomic Epidemiology [website].Lyngby: Technical University of Denmark. (http://www.genomicepidemiology.org/, accessed 10 April 2018).
23. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45(D1):D566-D73.
24. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis – 10 years on. Nucleic Acids Res. 2016;44(D1):D694-7.
25. Tsafnat G, Copty J, Partridge SR. RAC: Repository of antibiotic resistance cassettes. Database (Oxford). 2011;2011:bar054.
26. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42(Database issue):D581-91.
27. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 2017;45(D1):D535-D42.
28. PHIDIAS: Pathogen-Host Interaction Data Integration and Analysis System. Available from: http://wwwphidiasus/victors/indexphp.

29. Liu B, Pop M. ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009;37(suppl 1):D443–7.
30. Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R et al. PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. Nucleic Acids Res. 2017;45(D1):D604-D10.
31. Urban M, Irvine AG, Cuzick A, Hammond-Kosack KE. Using the pathogen-host interactions database (PHI-base) to investigate plant pathogen genomes and genes implicated in virulence. Front Plant Sci. 2015;6:605.
32. Urban M, Pant R, Raghunath A, Irvine AG, Pedro H, Hammond-Kosack KE. The Pathogen-Host Interactions database (PHI-base): additions and future developments. Nucleic Acids Res. 2015;43(Database issue):D645-55.
33. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Kohler J, Hammond-Kosack KE. PHI-base: a new database for pathogen host interactions. Nucleic Acids Res. 2006;34(Database issue):D459-64.
34. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. 2007;35(Database issue):D391-4.
35. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data. J Clin Microbiol. 2015;53(8):2410-26.
36. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH et al. The Salmonella In Silico Typing Resource (SISTR): an open Web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. PLoS One. 2016;11(1):e0147101.
37. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA et al. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015;53(5):1685-92.
38. Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M et al. Whole genome sequence-based serogrouping of Listeria monocytogenes isolates. J Biotechnol. 2016;235:181-6.
39. Deng X, den Bakker HC, Hendriksen RS. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. Annu Rev Food Sci Technol. 2016;7:353-74.
40. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinforma 2010. Chapter 11.
41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
42. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology. 1999; 99:45-56.
43. Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv.
44. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):524.
45. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728-36.
46. Instiut pasteur MLST databases and software [website]. Paris: Institut Pasteur. (http://bigsdb.pasteur.fr/., accessed 10 April 2018).
47. Rasko DA, Myers GS, Ravel J. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics. 2005;6:2.
48. BSR-Based Allele Calling Algorithm [website]. San Francisco: GitHub Inc. (https://github.com/mickaelsilva/chewBBACA., accessed 10 April 2018).
49. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis. 2016;63(3):380-6.
50. Petkau A MP, Sieffert C, Knox N, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. bioRxiv. 2016.
51. Rapid bacterial SNP calling and core genome alignments[website]. San Francisco: GitHub Inc. (https://github.com/tseemann/snippy)., accessed 10 April 2018).
52. «Reads to report» for public health and clinical microbiology[website]. San Francisco: GitHub Inc. (https://github.com/tseemann/nullarbor). NCBI, accessed 10 April 2018).
53. Pathogen Detection [website]. U.S. National Library of Medicine. Bethesda: National Center for Biotechnology Information. (ncbi.nlm.nih.gov/pathogens, accessed 10 April 2018)
54. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E.CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Science. 2015:1:e20 (https://doi.org/10.7717/peerj-cs.20, accessed 10 April 2018).
55. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol. 2016;54(8):1975-83.

56. Gardner SN, Hall BG. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One. 2013;8(12):e81760.

5. Use of WGS information by health professionals and risk managers: the need for cultural change

5.1 The role of microbiologists, bioinformaticians and epidemiologists

Turning the results from WGS into public health action requires close collaboration between (molecular) microbiologists, bioinformaticians and epidemiologists. The roles of each are detailed below, but it should be borne in mind that many of the tasks overlap and should be conducted collaboratively.

5.1.1 Molecularmicrobiologist

The molecular microbiologist is responsible for initiating and conducting sequencing. This includes:

- initial phenotypic and molecular identification and characterization of isolates, including culture purification and storage;
- genomic DNA extraction and purification, library preparation with appropriate quality controls;
- setting up of the sequencing run, which usually employs a high-throughput sequencing technology;
- downloading of sequencing data and review of quality measurements for the run;
- maintenance of accurate secure records of all procedures, including electronic databases of genome sequences and related data;
- appropriate record-keeping and accounting for and maintaining all equipment and consumables used;
- maintenance of culture collections of pathogens to allow retrospective audits and selection of internal control strains for WGS experiments;
- participation in quality control and quality assurance programmes to ensure national and international harmonization of sequencing methods;
- in collaboration with the bioinformatician, identification of reference genomes when required and determination of the requirements for an information technology (IT) environment that supports genomic data sharing, storage and archiving;
- in collaboration with the epidemiologist, determination of the data required to validate interpretation criteria and perform cluster assessments. This might include retrospective sequencing of isolates from well-defined outbreaks (i.e. with strong epidemiological evidence or an identified source);
- in collaboration with the epidemiologist, identify the challenges in implementation, processes to be used, and how and when isolates should be sequenced for optimal interpretation of surveillance information.

5.1.2 Bioinformatician

The bioinformatician is responsible for post-sequencing data analysis and storage of genomic data. (Routine data analyses may be performed by the molecular microbiologist or the e-lab technician.) The tasks include:

- computational analysis of sequencing data, usually in collaboration with microbiologists who are experts in the genomics of particular pathogens. The analysis may be done with off-the-shelf software, online tools or in-house or open-source pipelines;
- implementation, verification and management of computer-based algorithms for genome assembly, variant detection and isolate clustering through construction of phylogenetic trees;
- maintenance of accurate secure records of all procedures, including electronic databases of genome sequences and related quality control data;
- quality assessment of original and processed sequencing data;
- in collaboration with the microbiologist and epidemiologist, determination of the most appropriate method of analysis (e.g. de novo assembly, mapping to a reference);
- collaboration in the development of reports and cluster-naming conventions;
- in collaboration with the microbiologist and epidemiologist, assessment of clusters to determine which ones should be followed up and investigated.

5.1.3 Epidemiologist

The epidemiologist is responsible for collecting epidemiological information and integrating it with WGS data. This includes:

- in collaboration with the microbiologist and bioinformatician, determination of cluster nomenclature and the most appropriate method of reporting;
- setting of definitions for what constitutes a cluster (e.g. four *Salmonella* notifications in the previous two weeks with related genomic sequences) to support epidemiological investigations;
- determination of how WGS data will be implemented in the existing public health surveillance infrastructure and evaluation of the implementation to identify challenges, needs and gaps;
- determination of the format for reporting WGS outputs for routine surveillance and outbreak investigations (line lists only, line lists and trees, trees only, etc.);
- identification of database needs for incorporating the outputs from WGS into existing surveillance systems;
- combination of epidemiological information with reported WGS data;
- determination of which cases need to be followed up to collect epidemiological information, including determination of what isolates are part of the cluster (an outbreak may consist of more than one distinct genomic sequence);
- ensuring that appropriate information is collected to fulfil legal and legislative requirements;

- in conjunction with the microbiologist and bioinformatician, development and evaluation of data-sharing protocols that meet legal and legislative requirements; and
- monitoring of timelines of the WGS process to permit appropriate public health action.

The collaborative effort of epidemiologists, microbiologists and bioinformaticians relies on continuous communication. Table 5.1 outlines the skills of microbiologists, bioinformaticians and epidemiologists that contribute to the synthesis of epidemiological and genomic evidence for effective public health action.

TABLE. 5.1

Skills needed to translate WGS data into public health action

Bioinformatician	Epidemiologist	Microbiologist
Algorithms for genome mapping, assembly and comparisons	Epidemiology of communicable diseases	Microbiological diagnostics
Inferences from genomic data	Statistical analysis	Subtyping of pathogens
Genomic data handling and processing	Case-control studies	Pathogen genomics and evolution
Genome data visualization and integration	Health data linkage	Access to culture collections with epidemiological context
	Risk assessment and communication	

5.2 Integration of WGS, epidemiological, and clinical data

FBD surveillance requires continual analysis of laboratory data in order to identify potential outbreaks at the earliest possible stage. Information about the isolate must be integrated with the WGS result. This information, frequently referred to as metadata, describes the source and details of the isolate in terms of laboratory, epidemiological, and clinical characteristics (Table 5.2).

TABLE 5.2

Typical metadata associated with FBD surveillance

Type of metadata	Examples
Laboratory	Isolate, phenotypic results, QA/QC metrics, parameters of sequencing, assembly, and tree construction, dates (collection, analysis, reporting)
Epidemiological (clinical isolates)	Isolate source (clinical) isolate source details (specimen collection site), patient demographics (e.g. age, sex), exposures, geographical location
Food safety (non-clinical isolates)	Isolate source (food, environment), package type (intact, non-intact), isolate source details (food product type, geographical location), dates
Clinical	Patient immune status, underlying conditions, symptoms, dates (e.g. onset of illness), clinical outcome (e.g. hospitalization, death)

The collection of these metadata has long been the cornerstone of laboratory-based surveillance. Isolate information is necessary in order to make sense of laboratory results and inform public health action. Laboratory data should be stratified by person, place and time. For outbreak investigations in particular, analyses that integrate laboratory results with exposures, patient demographics and clinical features are critical in order to generate and test hypotheses about the source. To answer broader public health surveillance questions, such as those around risk assessment, source attribution, and ecosystems modelling, contextual information about the isolates is also required. Typically, laboratories have relied upon laboratory information management systems (LIMS) to manage metadata. These systems range from rudimentary handwritten notes or spreadsheets to sophisticated custom or commercial software packages that are fully integrated with laboratory equipment, and analysis and communication tools.

Genomics introduces new opportunities to use isolate information. Traditionally, FBD surveillance is analysed on a daily or weekly basis, in order to detect potential outbreaks. Once an outbreak is detected, laboratory results are analysed by exposure and molecular subtype, generating lists of cases. With WGS, the sequence data has sufficient resolution to confirm or exclude the epidemiological hypothesis made during an outbreak investigation. Thus, graphs representing the phylogenetic relationship between samples (e.gphylograms, cladograms, minimum spanning trees, split networks) is overlaid with metadata, bringing laboratory and epidemiological analyses together in real time. Bioinformatics tools focus on this critical need to integrate laboratory and epidemiological data and produce highly customizable visualizations. Specifically, Phyloviz Online has been designed as an open-source Web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees of allelic data extracted from WGS integrated with metadata *(1)*. Annotation of phylogenetic trees is also possible with interactive visualization Web-tools, such as iTOL*(2)* and Phandango*(3, 4)*. Similarly, a simple solution for data visualization and interpretation is provided by MicroReact, which allows users to upload, visualize and explore dendrograms linked to geographical location, time and other metadata *(5)*. Both Canada's IRIDA project and the United States' GenomeTrakr integrate laboratory data and epidemiological information to produce simple visual aids that can be understood and interpreted by public health laboratories and epidemiologists through an open source, end-to-end platform for infectious disease genomic epidemiology *(6-8)*. The commercial end-to-end platforms CLC Genomics Workbench (CLC Bio), BioNumerics (Applied Maths), and RidomSeqSphere+ (Ridom GmbH), which are widely used by public health laboratories, provide a similar level of integration of laboratory and epidemiological data.

5.3 Standardization of data and information and controlled vocabulary

A basic principle of data management is the standardization of data fields to ensure the data are sufficiently descriptive and well organized *(7)*. In the past, this could be managed via a local or regional database, with users implementing their own controlled vocabulary, aided perhaps by the use of drop-down menus. With WGS, the value of the data extends well beyond a single laboratory, or even a laboratory network; their application is potentially global, not only for immediate public health action, but also for long-term studies and applied research *(9)*. As a result, public health institutions are encouraged to make WGS data from routine surveillance activities publicly available, a practice that has been endorsed by PulseNet International *(10)* and GenomeTrakr*(6, 8)*. The vast datasets created can be used by anyone around the

world, greatly increasing the likelihood of semantic ambiguities and rendering the standardization of data vocabulary all the more critical. For example, different terms may exist to describe the same specimen type or food type ("beef", "ground beef", 'fresh ground beef", or "beef, fresh"). Without standardized terminology, this could be highly confusing. In line with global data management principles, data should be findable, accessible, interoperable, and reusable *(11)*. A critical step in achieving these standards is the development and implementation of an ontology, or controlled vocabulary, for all isolate information, laboratory analysis, clinical data, and epidemiological information. IRIDA's Genomic Epidemiology Application Ontology Consortium *(7)* maps community standards and existing ontologies to GenEpiO terms (Table 5.3). By mapping terms to a reference ontology, the impact of differences in vocabulary can be minimized; not only can the terms be easily understood, they are also machine-readable. This ensures that the meanings of terms are accurately maintained and improves interoperability between software systems *(7)*.

TABLE 5.3

Ontologies for use by public health laboratories

Ontology	Description	Link
OBO Foundry (*12*)	Development and collection of interoperable ontologies that are both logically well formed and scientifically accurate	http://www.obofoundry.org/
GenEpiO	Ontology for laboratory analytics, sample metadata, epidemiology, clinical data, and reporting	https://github.com/GenEpiO/genepio/wiki
FoodON	Farm-to-fork food ontology (food items, ingredients, production environments, risk assessment, source attribution	https://github.com/FoodOntology/foodon
TypOn(*13*)	Sequence-based microbial typing ontology (including WGS based methods)	https://bitbucket.org/phyloviz/typon

5.4 New paradigms of practice arising from developments in pathogen genomics

WGS-guided surveillance promises the rapid, precise identification of bacterial transmission pathways in hospital and community settings, with concomitant reductions in infections, morbidity and costs *(14-16)*. Because WGS offers unprecedented resolution for determining degrees of relatedness among bacterial and viral isolates, it complements existing epidemiological tools by allowing the reconstruction of transmission chains and identification of sequential acquisitions and otherwise unrecognized epidemiological links. For example, investigations of hospital outbreaks of methicillin-resistant *Staphylococcus aureus* and *Clostridium difficile* by WGS have allowed discrimination between apparently similar isolates collected within a short time-frame *(17, 18)*. In addition, recent studies have shown that WGS can detect super-spreaders, predict the existence of undiagnosed cases and intermediates in transmission chains, suggest likely direction of transmission, and identify unrecognized risk factors for onward transmission. These data are important in attempts to stop or minimise outbreaks, the design and evaluation of intervention programmes, and the allocation of public health resources.

Health professionals dealing with complex outbreaks, for which trace-back is complicated and labour-intensive, are greatly supported by genomics-enhanced surveillance with radically improved resolution. WGS has inspired a vision for "precision public health" with more effective public health actions and better patient outcomes. This new paradigm of practice is based on emerging statistical methods that can infer transmission networks and contact structures from pathogen genomic data, with or without contextual epidemiological information *(19-21)*. Genomics-based estimation of likely transmission pathways can greatly improve the tracking of transmission and our understanding of the mechanisms, as well as reducing our dependence on difficult-to-collect and often incomplete epidemiological data. These developments provide a better understanding of the epidemiology of high-burden infectious diseases and present new opportunities for proactive laboratory surveillance. Furthermore, genome sequences of local pathogens can easily be compared with other sequences in publicly available international databases, allowing the local outbreak to be interpreted in an international context, and often uncovering unexpected links to sources of infection elsewhere *(22)*. These opportunities are being explored in different countries. For example, the European Centre for Disease Prevention and Control has been running the ECDC Surveillance Systems Re-engineering Project *(23)*.

However, these new types of information require matching skills in public health professionals and cultural change in their practice. First, health professionals will need to acquire skills in multidimensional data analysis, public health genomics and evidence synthesis. Secondly, epidemiologists and microbiologists will need to play an increasingly important role in data governance and promotion of data sharing in an international context. There is a strong argument for including in the International Health Regulations a requirement that pathogen sequencing data should be shared. Lastly, the added value of WGS surveillance is maximized by breaking down the silos of epidemiology and microbiology and engaging informatics specialists in data analysis and visualization. At the same time, the increasingly recognized value of data-sharing creates new challenges of interdisciplinary communication and the communication of WGS results to the public.

Table 5.4 compares two approaches to conducting WGS in a public health laboratory. The first approach uses sequencing once an alert has been detected in routine surveillance data. WGS is used to confirm which cases belong in the outbreak and to assist in identifying the source of the outbreak. The second approach uses WGS for routine surveillance to detect small clusters.

TABLE 5.4

Two models of WGS-based surveillance

Variables	Epidemiological hypothesis-driven and WGS-guided outbreak investigations	Prospective, epidemiological, hypothesis-free, WGS-based laboratory surveillance
Type of pathogens	Any pathogen	High-burden, well-characterized bacterial or viral pathogens with established diagnostic and referral pathways
Examples of actionable information	Reconstruction of outbreak origins, transmission pathways and dating of transmission events; identification of fastidious pathogens at the species and lineage levels; source attribution and revealing the spatial spread of disease outbreak; identification of new clones associated with community- or hospital-acquired pathogens	Alerts about clusters according to predefined and validated rules; detection of outbreaks, covert clusters and associated risk factors; monitoring of endemic and sporadic activity; near real-time identification of transmission events; identification of new successful clones, transmission pathways and dating of transmission events within outbreaks (person-to-person contact; water- and foodborne modes of direct transmission)
Average size of outbreaks investigated	Usually large	Detection of outbreaks of all sizes, the majority are small
Required laboratory capacity	Moderate, WGS and bioinformatics analysis can be outsourced	High, in-house WGS and bioinformatics expertise
Complexity of interpretation of WGS result	Variable, depending on the size of the genome and the outbreak	High, especially for large-scale surveillance pathogens
Integration between public health and laboratory teams	Variable, depending on circumstances	Essential, via shared databases
Potential impact on health outcomes	Moderate to high	High, requires substantial additional resources for public health follow-up of surveillance alerts

5.5 References

1. Ribeiro-Goncalves B, Francisco AP, Vaz C, Ramirez M, Carrico JA. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res. 2016;44(W1):W246-51.
2. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44(W1):W242-5.
3. Hadfield, J.Phandango: Interactive visualization of genome phylogenies. (http://jameshadfield.github.io/phandango/ - /, accessed 27 February 2018).
4. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR.Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2017:34 (2):292–293.
5. Tool, Microreact- Hierarchical and Geographical Analysis. Available from: https://microreact.org/showcase.
6. NCBI. NCBI Pathogen Detection. Available from: https://www.ncbi.nlm.nih.gov/pathogens/, accessed 27 February 2018.
7. Canada, Public Health Agency of. IRIDA – Integrated Rapid Infectious Disease Analysis Project. Available from: http://www.irida.ca/, accessed 27 February 2018.
8. Nutrition, United States Food and Drug Administration's Center for Food Safety and Applied. GenomeTrakr Network. (https://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm, accessed 27 February 2018).
9. Food and Agriculture Organization of the United Nations, Applications of Whole Genome Sequencing in food safety management. (http://www.fao.org/3/a-i5619e.pdf, accessed 21 March 2018).
10. Nadon C., Van Walle, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International. Vision for the implementation of whole genome sequencing for global foodborne disease surveillance. Euro Surveill. 2017; 8:22(23):30544.
11. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A,et al. The FAIR Guiding Principles for scientific data management and stewardship.Sci Data, 2016;3:160018.
12. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. NatBiotechnol. 2007;25(11):1251-5.
13. Vaz C, Francisco AP, Silva M, Jolley KA, Bray JE, Pouseele H et al. TypOn: the microbial typing ontology. J Biomed Semantics. 2014;5(1):43.
14. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis. 2016;63(3):380-6.
15. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ. 2015;350:h1314.
16. Public health in anage of genomics. Paris: Organisation for Economic Co-operation and Development;2013 (OECD Science, Technology and Industry POlicy Papers, No. 8).
17. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. N Engl J Med. 2013;369(13):1195-205.
18. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469-74.
19. Gonzalez-Candelas F, Bracho MA, Wrobel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol. 2013;11:76.
20. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends in Microbiol. 2014;22(5):282-91.
21. Lipkin WI. The changing face of pathogen discovery and surveillance. Nat Rev Microbiol. 2013;11(2):133-41.
22. 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.
23. European Centre for Disease Prevention and Control. Technical Report: ECDC roadmap for integration of molecular typing into European-level surveillance and epidemic preparedness. Version 2.1, 2016-2019. Stockholm: ECDC; 2016.

ISBN 978 924 1 51386 9

World Health Organization