Clinical Microbiology and Infection xxx (2018) 1–5

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Review

Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches

A.C. Schürch [1], S. Arredondo-Alonso [1], R.J.L. Willems [1], R.V. Goering [2, *]

[1)] *Department of Medical Microbiology, University Medical Center, Utrecht, The Netherlands*
[2)] *Department of Medical Microbiology and Immunology, Creighton University School of Medicine, Omaha, NE, USA*

ARTICLE INFO

Article history:
Received 6 July 2017
Received in revised form
21 November 2017
Accepted 22 December 2017
Available online xxx

Editor: P.T. Tassios

Keywords:
Accessory genome
cgMLST
Core genome
Pan-genome
Relatedness threshold
SNP
Whole genome sequencing

ABSTRACT

Background: Whole genome sequence (WGS)-based strain typing finds increasing use in the epidemiologic analysis of bacterial pathogens in both public health as well as more localized infection control settings.
Aims: This minireview describes methodologic approaches that have been explored for WGS-based epidemiologic analysis and considers the challenges and pitfalls of data interpretation.
Sources: Personal collection of relevant publications.
Content: When applying WGS to study the molecular epidemiology of bacterial pathogens, genomic variability between strains is translated into measures of distance by determining single nucleotide polymorphisms in core genome alignments or by indexing allelic variation in hundreds to thousands of core genes, assigning types to unique allelic profiles. Interpreting isolate relatedness from these distances is highly organism specific, and attempts to establish species-specific cutoffs are unlikely to be generally applicable. In cases where single nucleotide polymorphism or core gene typing do not provide the resolution necessary for accurate assessment of the epidemiology of bacterial pathogens, inclusion of accessory gene or plasmid sequences may provide the additional required discrimination.
Implications: As with all epidemiologic analysis, realizing the full potential of the revolutionary advances in WGS-based approaches requires understanding and dealing with issues related to the fundamental steps of data generation and interpretation. **A.C. Schürch, Clin Microbiol Infect 2018;▪:1**

What is the biologic basis of diversity in bacterial genomes?

Bacterial genomes of the same species contain a common set of genes referred to as the core genome. However, variability between bacterial genomes of the same species occurs for a variety of reasons, including point mutations, homologous recombination and differences in genome content [1–3]. Point mutations encompass single nucleotide polymorphisms (SNPs) and single nucleotide insertions or deletions at variable mutation rates that can range widely depending on the species. For example, while *Mycobacterium tuberculosis* may accumulate approximately four SNPs per genome per 4 years [4], *Helicobacter pylori* can exhibit over 30 SNPs per genome per year [5]. Recombination, the exchange of genetic material between bacteria, ranges from rare (*M. tuberculosis*) [6] to strong enough for large-scale confusion of phylogenetic interrelationships in gene trees (*Streptococcus pneumoniae, H. pylori*) [7,8]. The causes of variability in genome content are large insertions, deletions, genome rearrangements and transfer of exogenous DNA, and extrachromosomal elements such as plasmids and phages. This results in a set of genes that is variably present in sequenced members of a species and is referred to as the accessory gene content or accessory genome. Taken together, the pangenome represents all genes, whether constant or variable, that are found in members of a species [9,10].

There are a variety of approaches to analysing these genomic data for epidemiologic and infection control purposes. At a high level, phylotyping (strain interrelationships based on sequence-associated evolutionary history) may support epidemiologic

* Corresponding author. R. V. Goering, Department of Medical Microbiology and Immunology, Creighton University School of Medicine, 2500 California Plaza, Omaha, NE 68178, USA.
E-mail address: richardgoering@creighton.edu (R.V. Goering).

https://doi.org/10.1016/j.cmi.2017.12.016

investigations to determine the source and routes of infections, trace cross-transmission of healthcare-associated pathogens and identify virulent antibiotic-resistant lineages or subpopulations. This subject is discussed in more depth elsewhere in this issue, while more sensitive SNP and gene-by-gene typing approaches indexing core or accessory genome variation due to mutations or recombination events are considered here.

How can SNP-based approaches be performed?

Strain typing by whole genome sequencing (WGS) based on SNPs can be performed via reference-based mapping of either reads or assembled contigs. Many studies choose to map sequencing reads against a reference genome [11–13] with either a custom pipeline or using one of the available microbial SNP pipelines, such as Snippy, NASP, SNVphyl, CFSAN SNP Pipeline, or Lyve-SET (https://github.com/tseemann/snippy) [14–16]. There are many other similar tools, but in the context of this review, we will name only one or two examples of software solutions for this discussion and that of core genome alignment that follows. Another review in this series gives further detail and insight into additional available bioinformatic tools [70].

How can core genome alignment be performed?

Another method to identify SNPs is to build a core genome alignment. By identifying a core genome as sets of orthologous (i.e. common ancestor) sequences conserved in all aligned genomes, it is possible to focus either on identification of groups of orthologous genes (Roary is one of the most recent pan-genome analysis tools [17]) or on collinear blocks of orthology, which includes intergenic regions that might break blocks within genes (e.g. Parsnp implemented in the Harvest suite) [18]. Irrespective of the method used, a core genome alignment might contain stretches of high variability, potentially resulting from recombination events. Including this variability in the phylogenetic tree can interfere with the phylogenetic signal. In a species-specific tool such as Whole Genome Sequence Analysis (http://www.wgsa.net/) (currently for *Staphylococcus aureus, Salmonella typhi, Neisseria gonorrhoeae, Renibacterium salmoninarum,* and Zika virus), these stretches are excluded before tree building by mapping assembled genomes against a curated collection of species reference genomes. However, sequence diversity, potentially the result of recombination, needs to be dealt with, usually by identification and filtering of SNPs with specialized tools (e.g. Gubbins, PhiPack and BRATNextGen) [18–20].

The SNP approach can give very high resolution, but a reference genome must be used that is closely related to the sequenced samples. This reduces the chances of mismapping and increases the regions present in the reference genome to which reads will be mapped against. The chance of mismapping increases and the number of mapped bases decreases if a diverse set of samples is mapped against an arbitrary reference. Ideally, the analysed samples are also very closely related, which is the case in outbreak or intrapatient divergence studies. In this setting, one of a set of closely related samples may be assembled and used as reference against which all others can be mapped [21,22]. The resulting alignment is then the basis for phylogenetic reconstruction, either directly or after concatenation of the positions showing variance. Another drawback of the SNP-based approach is the low comparability between different studies, especially if different reference genomes are used, and the potential adoption of different threshold settings, such as the parameters by which a SNP will be verified by the analysis software. Thus, reproducibility of a specific SNP analyses requires that the pipeline, its settings, the reference genome and reads for a particular study are publicly available. In this context, an effort to establish benchmark data sets for phylogenetic pipelines has been recently proposed that will allow validation and comparison of analyses in specific settings such as food-borne pathogen outbreak surveillance [23].

How can gene-by-gene comparison be performed?

An alternative approach to analyse genetic relatedness is a modification of traditional multilocus sequence typing (MLST) described by Maiden et al. [24], termed whole genome (wg)/core genome (cg) MLST (or extended MLST) (Table 1). Here, genome-wide gene-by-gene comparisons of hundreds to more than a thousand genes allow the assignment of alleles in comparison to a curated set of predefined core genes, which ensures interlaboratory reproducibility. Public schemes have been developed and are maintained for many, but not all, important pathogens. Analyses can be performed with a variety of different software, including BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium), SeqSphere+ (Ridom, Münster, Germany), and BIGSdb (https://pubmlst.org/software/database/bigsdb/). For species without published schemes, or to include as many loci as possible to further improve resolution, *ad hoc* schemes can be developed. However, an inherent danger is concurrent development of divergent approaches, as is currently the case for *Klebsiella pneumonia* (http://www.cgmlst.org/ncs) [29,52,53]. The common typing language inherent in cgMLST can be used to support analysis in local infection control [31,54] as well as a broader public health setting [55]. Furthermore, recombination is handled differently by cgMLST than by a SNP alignment, as recombinant regions with a high density of SNPs can be filtered, while cgMLST methods will collapse these regions into a smaller number of allelic changes.

How is accessory genome typing performed?

Both SNP and cgMLST analyses are restricted to regions of the genome present in all analysed isolates, which means that potentially useful information in the accessory genome is discarded. Recently the pan-genome has been advocated and applied as an additional tool to type bacterial genomes [56,57]. Clustering of strains based on the presence or absence of accessory genes allows a sensitive, fine-grained analysis of isolates which will not necessarily be concordant with core genome analysis. However, combined analysis of variation in core, accessory and regulatory genome regions can provide a superresolution view into the epidemiology of bacterial populations [57]. Given the proposed function of the pan-genome in adaptive evolution [10], this approach could also give better insight in niche specification than cgMLST or SNP typing alone. Therefore, wider application of accessory genome typing is anticipated in the near future.

How can plasmids be reconstructed from WGS data?

Despite the importance of plasmids (e.g. in the transmission of antimicrobial resistance genes), surveillance and outbreaks analyses are frequently focused on tracing clonal strains through SNPs or cgMLST. This strategy may not be suitable in scenarios where plasmid conjugation or transposition are highly frequent, leading to a plasmid-mediated rather than clonal outbreak [59,58]. To fully understand antimicrobial resistance introduction and transmission in hospital environments, plasmid epidemiology must be taken into account [60,61].

Low-level resolution obtained by PCR-based techniques and laborious work associated with plasmid purification are the main obstacles to using these techniques to analyse large collections of

Please cite this article in press as: Schürch AC, et al., Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches, Clinical Microbiology and Infection (2018), https://doi.org/10.1016/j.cmi.2017.12.016

Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

Organism	Relatedness threshold[a]		References
	wg/cgMLST (allele) SNPs		
Acinetobacter baumannii	≤8	≤3	[25,26]
Brucella spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Campylobacter coli, C. jejuni	≤14	≤15	[27,28]
Cronobacter spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Clostridium difficile	Epidemiologic validation in progress[b]	≤4	[29], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst
Enterococcus faecium	≤20	≤16	[30]
Enterococcus raffinosus	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Escherichia coli	≤10	≤10	[31,32], https://enterobase.warwick.ac.uk/
Francisella tularensis	≤1	≤2	[33,34]
Klebsiella oxytoca	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Klebsiella pneumonia	≤10	≤18	[35,36]
Legionella pneumophila	≤4	≤15	[37]
Listeria monocytogenes	≤10	≤3	[38,39]
Mycobacterium abscessus		≤30	[40]
Mycobacterium tuberculosis	≤12	≤12	[41]
Neisseria gonorrhoeae	Epidemiologic validation in progress[b]	≤14	[42], http://www.applied-maths.com/applications/wgmlst
Neisseria meningitidis	Epidemiologic validation in progress[b]		http://www.cgmlst.org/ncs
Pseudomonas aeruginosa	≤14	≤37	[31,43]
Salmonella dublin	Epidemiologic validation in progress[b]	≤13	[44], https://enterobase.warwick.ac.uk/
Salmonella enterica	Epidemiologic validation in progress[b]	≤4	[45], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/
Salmonella typhimurium	Epidemiologic validation in progress[b]	≤2	[46], https://enterobase.warwick.ac.uk/
Staphylococcus aureus	≤24	≤15	[47,48]
Streptococcus suis		≤21	[49]
Vibrio parahaemolyticus	≤10		[50]
Yersinia spp.	0		[51]

cg, core genome; MLST, multilocus sequence typing; SNP, single nucleotide polymorphism; wg, whole genome.
[a] Data often represent single studies that can be used to begin formulation of species-specific interpretation criteria. Thus, these data should be coupled with newly published similar studies to ensure that resulting values are not atypical and can be generally applied.
[b] Proposed wg/cgMLST schemes are available online (http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/) but as yet have not been epidemiologically validated.

isolates. Accordingly, WGS has been adopted as the reference standard to analyse plasmid sequences [62,63]. However, short reads (read length up to 300 bp) generated by high-throughput sequencers (e.g. Illumina MiSeq) cannot span plasmid repeat sequences, leading to an accurate but fragmented assembly. Several tools had been proposed to improve *de novo* plasmid assembly, but manual and expert pruning is required to obtain correct plasmid boundaries, which limits the high-throughput analysis of WGS data [64]. Emerging long-read sequencing technologies are a promising solution to obtain complete and error-free plasmid sequences [65]; they were used to resolve the complete sequence of a 165 kbp plasmid sequence, confirming an alternative transfer mechanism of the yersiniabactin locus *ybt* in *K. pneumonia* (http://www.biorxiv.org/content/early/2017/01/04/098178) [71].

Data interpretation and issues of isolate relatedness

As stated previously, for purposes of bacterial strain typing and epidemiologic analysis, WGS is primarily managed by SNP and/or gene-by-gene (e.g. cgMLST) comparisons. While both typing approaches may in some cases provide comparable epidemiologic discrimination [66], they vary in their overall suitability for different epidemiologic or infection control settings. At present, strain typing for public health is often performed by cgMLST, with its capability for readily scaling analyses from small to large isolate numbers, national and international data sharing, and standardized strain nomenclature [55]. As noted earlier, the establishment of curated cgMLST databases and commercially available software with graphical user interfaces may reduce the need for dedicated bioinformatics expertise. In addition, the National Center for Biotechnology Information (NCBI) has developed the NCBI Pathogen Detection Portal (https://www.ncbi.nlm.nih.gov/pathogens/), a robust platform using a combination of kmer and SNP methods that has analysed more than 150 000 isolates from 20 different species. This resource is available to users willing to make their data public at NCBI, with no bioinformatic expertise required for data analysis. Smaller-scale strain typing for infection control may be less reliant on many of the above analytical resources and frequently benefits from the increased sensitivity of SNP analysis. However, a SNP-based network for tracking food-borne pathogens with WGS is currently in place [67].

As with all typing methods, WGS analysis involves actively growing isolates where genomic variability (i.e. clock speed) directly affects the ultimate strain typing issue of relatedness, which increases in complexity over time [68]. In the frequent absence of an index isolate, the challenge is to determine what constitutes significant similarity and/or difference among the isolates available for analysis. While past typing methods such as pulsed-field gel electrophoresis lent themselves to generalized guidelines for assessing isolate interrelationships [69], the volume of WGS data and its potential organism-specific variability over time resists the establishment of general relatedness guidelines. Thus, there are no easy thresholds of relatedness, and the issue of significant difference must be established on an organism-by-organism and case-by-case basis. Knowledge of organism-specific population genetics (i.e. the relative impact of mutation and recombination on genetic variation) is crucial to correctly interpret genetic differences among strains. Well-conducted studies of isolate populations that include known epidemiologically linked as well as unrelated isolates with associated clinical metadata are also vital to the establishment of similarity benchmarks. An example of current suggested cgMLST and SNP relatedness criteria for

Please cite this article in press as: Schürch AC, et al., Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches, Clinical Microbiology and Infection (2018), https://doi.org/10.1016/j.cmi.2017.12.016

representative clinically relevant bacteria is shown in Table 1. We are currently in the discovery phase of this process, which will improve both in quantity and quality over time. However, it is important to emphasize that proposed thresholds of clonality must always be considered as guidelines rather than absolute rules and thus be applied with a degree of flexibility. As with any family tree, diversity within related members is directly influenced by the number of represented generations. The longer the analysis timeline, the greater the possibility that a comparison of the most recent members (isolates) will exceed a suggested relatedness threshold (Fig. 1). In this context, a broad graphical overview of isolate interrelationships (e.g. the branching of a phylogenetic tree) generated by software such as that mentioned above can greatly inform epidemiologic interpretation. In addition, isolates close to (but beyond) the relatedness threshold, which could represent more distant members of an outbreak, can be targeted for further scrutiny.

The way forward

Proper interpretation of strain typing data has always underscored the need to consider epidemiologic concordance. In this context, past issues with the discriminatory power of available molecular typing methods have generally led to the conclusion that one can only exclude strains from an outbreak (e.g. label them as epidemiologically unrelated when diversity passes a preset threshold) but not include strains solely on the basis of typing without additional clinical data. The exquisite sensitivity of WGS-based typing causes us to consider revisiting this thinking. For example, one could speculate that, in the absence of clinical data, isolates of *Enterococcus faecium* (a highly diverse species) differing by one or two SNPs but from unrelated, geographically distinct hospitals could represent a 'novel' transmission route not as yet recognized by clinical epidemiologists. While this and other issues remain to be clarified, what is certain is that WGS has introduced a paradigm shift in the acquisition, analysis and interpretation of epidemiologic data. An exciting future lies ahead as these and other as-yet-unforeseen issues are identified and dealt with as revolutionary advances in WGS-based epidemiologic analysis continue to unfold.

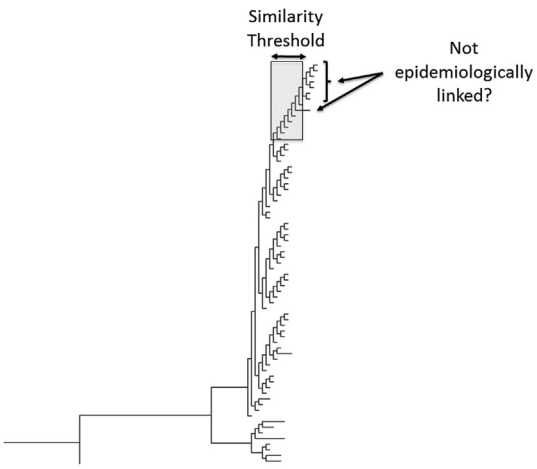

Fig. 1. Hypothetical dendrogram (neighbour-joining tree). Interpretation of whole genome sequencing results for epidemiologic investigations should rely on both genetic distance/difference (single nucleotide polymorphisms or alleles) and the topology of the dendrogram/phylogeny to draw conclusions about relatedness. Although a similarity threshold can act as a guide to identify clusters of potentially related isolates (e.g. within boxed area), isolates beyond the threshold but topologically nearby deserve further scrutiny for potential relatedness.

Transparency declaration

All authors report no conflicts of interest relevant to this minireview.

References

[1] Bryant J, Chewapreecha C, Bentley SD. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. Future Microbiol 2012;7:1283–96.
[2] Levin BR, Bergstrom CT. Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. Proc Natl Acad Sci U S A 2000;97:6981–5.
[3] Darmon E, Leach DR. Bacterial genome instability. Microbiol Mol Biol Rev 2014;78:1–39.
[4] Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. PLoS Med 2016;13:e1002137.
[5] Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, et al. *Helicobacter pylori* genome evolution during human infection. Proc Natl Acad Sci U S A 2011;108:5033–8.
[6] Liu X, Gutacker MM, Musser JM, Fu YX. Evidence for recombination in *Mycobacterium tuberculosis*. J Bacteriol 2006;188:8169–77.
[7] Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 2001;98:182–7.
[8] Cao Q, Didelot X, Wu Z, Li Z, He L, Li Y, et al. Progressive genomic convergence of two *Helicobacter pylori* strains during mixed infection of a patient with chronic gastritis. Gut 2015;64:554–61.
[9] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. Proc Natl Acad Sci U S A 2005;102:13950–5.
[10] McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol 2017;2:17040.
[11] Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 2013;13:110.
[12] Reuter S, Ellington MJ, Cartwright EJ, Koser CU, Torok ME, Gouliouris T, et al. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med 2013;173:1397–404.
[13] 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;54:333–42.
[14] Sahl JW, Lemmer D, Travis J, Schupp JM, Gillece JD, Aziz M, et al. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microb Genom 2016;2, e000074.
[15] Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb Genom 2017;3:e000116.
[16] Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, et al. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. Front Microbiol 2017;8:375.
[17] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3.
[18] 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;15:524.
[19] Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 2012;40:e6.
[20] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 2015;43:e15.
[21] Haley BJ, Kim SW, Pettengill J, Luo Y, Karns JS, Van Kessel JA. Genomic and evolutionary analysis of two *Salmonella enterica* serovar Kentucky sequence types isolated from bovine and poultry sources in North America. PLoS One 2016;11:e0161225.
[22] van der Graaf-van Bloois Duim B, Miller WG, Forbes KJ, Wagenaar JA, Zomer A. Whole genome sequence analysis indicates recent diversification of mammal-associated *Campylobacter fetus* and implicates a genetic factor associated with H_2S production. BMC Genomics 2016;17:713.
[23] Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ 2017;5:e3893.
[24] Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 1998;95:3140–5.
[25] Higgins PG, Prior K, Harmsen D, Seifert H. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence–based typing of *Acinetobacter baumannii*. PLoS One 2017;12:e0179228.

Please cite this article in press as: Schürch AC, et al., Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches, Clinical Microbiology and Infection (2018), https://doi.org/10.1016/j.cmi.2017.12.016

[26] Halachev MR, Chan JZ, Constantinidou CI, Cumley N, Bradley C, Smith-Banks M, et al. Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant *Acinetobacter baumannii* in Birmingham, England. Genome Med 2014;6:70.
[27] Cody AJ, McCarthy ND, Jansen van RM, Isinkaye T, Bentley SD, Parkhill J, et al. Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 2013;51:2526–34.
[28] Llarena AK, Taboada E, Rossi M. Whole-genome sequencing in epidemiology of *Campylobacter jejuni* infections. J Clin Microbiol 2017;55:1269–75.
[29] Kumar N, Miyajima F, He M, Roberts P, Swale A, Ellison L, et al. Genome-based infection tracking reveals dynamics of *Clostridium difficile* transmission and disease recurrence. Clin Infect Dis 2016;62:746–52.
[30] de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van SW, et al. Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 2015;53:3788–97.
[31] Dekker JP, Frank KM. Next-generation epidemiology: using real-time core genome multilocus sequence typing to support infection control policy. J Clin Microbiol 2016;54:2850–3.
[32] Roer L, Hansen F, Thomsen MC, Knudsen JD, Hansen DS, Wang M, et al. WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. J Antimicrob Chemother 2017;72:1922–9.
[33] Afset JE, Larssen KW, Bergh K, Larkeryd A, Sjodin A, Johansson A, et al. Phylogeographical pattern of *Francisella tularensis* in a nationwide outbreak of tularaemia in Norway, 2011. Euro Surveill 2015;20:9–14.
[34] Antwerpen MH, Prior K, Mellmann A, Hoppner S, Splettstoesser WD, Harmsen D. Rapid high resolution genotyping of *Francisella tularensis* by whole genome sequence comparison of annotated genes ('MLST+'). PLoS One 2015;10:e0123298.
[35] Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 2012;4:148ra116.
[36] Zhou H, Liu W, Qin T, Liu C, Ren H. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Klebsiella pneumoniae*. Front Microbiol 2017;8:371.
[37] David S, Mentasti M, Tewolde R, Aslett M, Harris SR, Afshar B, et al. Evaluation of an optimal epidemiological typing scheme for *Legionella pneumophila* with whole-genome sequence data using validation guidelines. J Clin Microbiol 2016;54:2135–48.
[38] 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;53:2869–76.
[39] Kvistholm JA, Nielsen EM, Bjorkman JT, Jensen T, Muller 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;63:64–70.
[40] Trovato A, Baldan R, Costa D, Simonetti TM, Cirillo DM, Tortoli E. Molecular typing of *Mycobacterium abscessus* isolated from cystic fibrosis patients. Int J Mycobacteriol 2017;6:138–41.
[41] 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 Clin Microbiol 2014;52:2479–86.
[42] De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, et al. Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. Lancet Infect Dis 2016;16:1295–303.
[43] Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, et al. Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. Euro Surveill 2013;18.
[44] Agren EC, Wahlstrom H, Vesterlund-Carlson C, Lahti E, Melin L, Soderlund R. Comparison of whole genome sequencing typing results and epidemiological contact information from outbreaks of Salmonella Dublin in Swedish cattle herds. Infect Ecol Epidemiol 2016;6:31782.
[45] Bekal S, Berry C, Reimer AR, Van DG, Beaudry G, Fournier E, et al. Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. J Clin Microbiol 2016;54:289–95.
[46] Phillips A, Sotomayor C, Wang Q, Holmes N, Furlong C, Ward K, et al. Whole genome sequencing of *Salmonella* Typhimurium illuminates distinct outbreaks caused by an endemic multi-locus variable number tandem repeat analysis type in Australia, 2014. BMC Microbiol 2016;16:211.
[47] Bartels MD, Larner-Svensson H, Meiniche H, Kristoffersen K, Schonning K, Nielsen JB, et al. Monitoring meticillin resistant *Staphylococcus aureus* and its spread in Copenhagen, Denmark, 2013, through routine whole genome sequencing. Euro Surveill 2015;20.
[48] Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol 2014;52:2365–70.
[49] Du P, Zheng H, Zhou J, Lan R, Ye C, Jing H, et al. Detection of multiple parallel transmission outbreak of *Streptococcus suis* human infection by use of genome epidemiology, China, 2005. Emerg Infect Dis 2017;23:204–11.
[50] Gonzalez-Escalona N, Jolley KA, Reed E, Martinez-Urtaza J. Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*. J Clin Microbiol 2017;55:1682–97.
[51] Kingry LC, Rowe LA, Respicio-Kingry LB, Beard CB, Schriefer ME, Petersen JM. Whole genome multilocus sequence typing as an epidemiologic tool for *Yersinia pestis*. Diagn Microbiol Infect Dis 2016;84:275–80.
[52] Wyres KL, Holt KE. *Klebsiella pneumoniae* population genomics and antimicrobial-resistant clones. Trends Microbiol 2016;24:944–56.
[53] Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, et al. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerg Infect Dis 2014;20:1812–20.
[54] Mellmann A, Bletz S, Boking T, Kipp F, Becker K, Schultes A, et al. Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting. J Clin Microbiol 2016;54:2874–81.
[55] Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017;22:30544.
[56] Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbe New Infect 2015;7:72–85.
[57] McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 2016;12:e1006280.
[58] Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, et al. Nested Russian doll–like genetic mobility drives rapid dissemination of the carbapenem resistance gene bla_{KPC}. Antimicrob Agents Chemother 2016;60:3767–78.
[59] Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 2016;16:161–8.
[60] 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;6:254ra126.
[61] Conlan S, Park M, Deming C, Thomas PJ, Young AC, Coleman H, et al. Plasmid dynamics in kpc-positive *Klebsiella pneumoniae* during long-term patient colonization. MBio 2016;7.
[62] Brolund A, Franzen O, Melefors O, Tegmark-Wisell K, Sandegren L. Plasmidome-analysis of ESBL-producing *Escherichia coli* using conventional typing and high-throughput sequencing. PLoS One 2013;8:e65793.
[63] de Toro M, Garcilláon-Barcia MP, De La Cruz F. Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. Microbiol Spectr 2014;2.
[64] Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, et al. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. Front Microbiol 2017;8:182.
[65] Judge K, Hunt M, Reuter S, Tracey A, Quail MA, Parkhill J, et al. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. Microb Genom 2016;2:e000085.
[66] Cunningham SA, Chia N, Jeraldo PR, Quest DJ, Johnson JA, Boxrud DJ, et al. Comparison of whole-genome sequencing methods for analysis of three methicillin-resistant *Staphylococcus aureus* outbreaks. J Clin Microbiol 2017;55:1946–53.
[67] 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;54:1975–83.
[68] Goering RV, Kock R, Grundmann H, Werner G, Friedrich AW. From theory to practice: molecular strain typing for the clinical and public health setting. Euro Surveill 2013;18:20383.
[69] Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 1995;33:2233–9.
[70] Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A Primer on Microbial Bioinformatics for non-bioinformaticians. J Clin Microbiol 2018. In press.
[71] Lam MCM, Wick RR, Wyres KL, Gorrie C, Judd LM, et al. Frequent emergence of pathogenic lineages of Klebsiella pneumoniae via mobilisation of yersiniabactin and colibactin. bioRxiv 2018:098178. https://doi.org/10.1101/098178.

Please cite this article in press as: Schürch AC, et al., Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches, Clinical Microbiology and Infection (2018), https://doi.org/10.1016/j.cmi.2017.12.016