Published in final edited form as:

Clin Chem. 2014 April ; 60(4): 586–588. doi:10.1373/clinchem.2013.217141.

What's Your Subtype? The Epidemiologic Utility of Bacterial Whole-Genome Sequencing

Thomas J. Sandora[1], **Peter Gerner-Smidt**[2], and **Alexander J. McAdam**[3,*]

[1]Division of Infectious Diseases, Departments of Medicine and Laboratory Medicine, Boston Children's Hospital, Boston MA

[2]Enteric Diseases Laboratory Branch, Centers for Disease Control & Prevention, Atlanta, GA

[3]Department of Laboratory Medicine, Boston Children's Hospital, Boston, MA

Keywords

Bacterial subtyping; whole genome sequencing

Determining the source of an infection can be difficult, but it is important for guiding interventions to interrupt outbreaks and prevent infections. Although many pathogens spread by person-to-person transmission, infections can also be acquired from food, animal, insect or environmental sources. A publication by Eyre et al. in the New England Journal of Medicine sheds light on the source of infection by the bacterium *Clostridium difficile* by use of a relatively new method of bacterial strain subtyping (1). Bacterial whole genome sequencing (WGS) is an emerging, broadly applicable and powerful technology that has the potential soon to replace multiple functions in research, clinical and public health microbiology laboratories.

Bacterial subtyping determines the similarity between separate isolates of bacteria of the same species. If bacteria have the same subtype, they are more likely to be related to each other than if they have different subtypes. Subtyping is used in epidemiologic investigations to gather information about microbial transmission. For example, if bacteria isolated from two patients sharing a hospital room are of the same subtype, the patients might have become infected from a common source or the infection may have been transmitted directly or indirectly from one patient to the other, but it is less likely that they acquired the infection from unrelated sources.

WGS is one of several available methods for bacterial subtyping. WGS generally has greater discriminatory power than other subtyping methods. The discriminatory power of a subtyping method is its ability to differentiate between epidemiologically unrelated strains of bacteria. Many genotypic methods for bacterial subtyping are available with varying discriminatory power (Table). Until recently, molecular subtyping methods utilized only a

*Address correspondence to this author at: Boston Children's Hospital, 300 Longwood Ave., Boston, MA 02115. Fax 617-730-0383; alexander.mcadam@childrens.harvard.edu.

This manuscript has not been previously presented.

small fraction of the three to five million bases in the genomes of most bacterial pathogens. For example, pulsed field gel electrophoresis (PFGE), a restriction fragment length polymorphism based method, remains the gold standard for highly discriminatory subtyping although it utilizes a tiny fraction of the genome: between 10 and 40 occurrences of a specific four to eight base sequence. Another less discriminatory subtyping method is multilocus sequence typing (MLST), which detects variations in the sequences of short regions (350 to 600 base pairs) in two to ten carefully selected genes.

Eyre et al. used bacterial WGS to investigate what proportion of *Clostridium difficile* infections (CDI) could be attributed to transmission from symptomatic patients (1). *C. difficile* causes a range of clinical manifestations, from asymptomatic intestinal carriage to fulminant or fatal pseudomembranous colitis. The organism has traditionally been thought to be acquired primarily by person-to-person spread through the fecal-oral route, although environmental sources also play a role (2). The recommended practices to prevent CDI include use of contact precautions and private patient rooms for infected patients and attention to environmental disinfection.

To investigate sources of CDI, Eyre et al. performed WGS on *C. difficile* isolates from patients diagnosed with symptomatic CDI at four hospitals that provide all acute care and 90% of hospital services in Oxfordshire, UK. WGS subtyping results from patient isolates were used to link cases; if two patient isolates matched, it was presumed that person-to-person transmission had occurred from a patient with CDI. The study period was long (three and a half years, although isolates obtained in the first six months were analyzed only as potential sources of infection), and the number of isolates tested was large (1,223). Epidemiologic data were collected to investigate links between the patients consistent with a transmission event (e.g. admission to the same hospital ward in a relevant timeframe). The WGS data were interpreted by mapping them to a reference genome and then performing pairwise comparisons between the sequenced isolates to identify single nucleotide variants (SNVs). This method provides robust data, but it is labor intensive and might not be ideal when results are needed quickly, e.g. in the clinical setting.

Before the investigators could interpret the WGS results for epidemiologic use, they had to define how much genetic variability defines a subtype of *C. difficile*. Bacteria divide rapidly and SNVs accumulate as the DNA is copied. The rate at which SNVs accumulate needs to be estimated to interpret WGS subtyping results. Eyre et al. used paired first and last isolates of *C. difficile* from patients who had multiple positive samples to estimate an evolutionary rate of 0.74 SNVs per year. They calculated a 95% prediction interval that 0-2 and 0-3 SNVs would be expected for isolates detected <124 and 124–364 days apart respectively. They used 0-2 SNVs to define isolates as genetically related, regardless of the length of time separating the detection of the isolates.

The key finding of this study was that a minority of CDI could be attributed to transmission from symptomatic patients, suggesting that that there are likely to be multiple important sources of *C. difficile* infection. Only 35% of patient isolates were genetically related to one or more isolates collected earlier in the study. It is striking that such a small proportion of CDI isolates were genetically linked, given previous hypotheses about hospital transmission.

A total of 45% of isolates had more than 10 SNVs compared with all previous isolates in the study, suggesting that patients with these isolates were infected from some source other than patients with CDI. Sources might include asymptomatically colonized individuals or environmental reservoirs.

The investigators compared subtyping results obtained by WGS and MLST. Among pairs of patients whose isolates were of the same MLST subtype and who were simultaneously on the same hospital ward when infection might have occurred, 28% of isolate pairs had more than 10 SNVs separating them. Among MLST-subtype matched patient pairs with more remote hospital contact, fully 76% of isolate pairs had more than 10 SNVs different. It is not surprising that WGS had much greater discriminatory power than MLST, since MLST only queries a small fraction of the genome whereas WGS may interrogate the whole genome.

Although most research using bacterial WGS on clinical samples has been for outbreak investigations, the method has been used for other applications that illustrate its potential to perform many functions in the microbiology laboratory. For example, WGS was used to find new genetic markers and confirm the importance of known genes associated with antibiotic resistance in *Mycobacterium tuberculosis* (3). It is not known whether the results of WGS can predict the results of functional antibiotic susceptibility tests with certainty since not all genes are expressed; if a new resistance mechanism arises, the genes encoding it will also need to be determined before WGS can detect it. WGS has also been used to investigate the origins and mechanisms of virulence of an extremely virulent strain of *Escherichia coli* serotype O104:H4 that caused an outbreak of hemolytic-uremic syndrome in Germany in 2011 (4). WGS results showed that this strain had characteristics of two pathotypes, enteroaggregative and Shiga toxin-producing *E. coli*. There are several different pathotypes of *E. coli*, each associated with specific clinical and epidemiologic characteristics, and WGS may replace cumbersome tests currently used to identify these pathotypes. WGS could also be an accurate method for determining the serotype of bacteria. Finally, WGS is the ultimate tool for public health surveillance of bacterial diseases to detect and investigate outbreaks. A collaborative effort is underway between the state public health laboratories, Centers for Disease Control & Prevention, United States Food and Drug Administration, United States Department of Agriculture and National Center for Biotechnology Information to detect and investigate outbreaks of *Listeria monocytogenes*. In this project, WGS is compared with the current gold standard, PFGE combined with intensive epidemiological follow-up. Multiple analytical approaches to WGS will be tested in order to establish which works best for this organism. It is expected that this project will prove high public health impact of WGS as an outbreak surveillance tool.

There are hurdles to be cleared before WGS is ready for routine use in clinical and public health laboratories. First, the cost of the sequencers is high and their operation requires significant technical expertise; however, the sequencers are rapidly becoming cheaper and simpler to operate. Second, the interpretation of WGS data is very complex and requires specific expertise, special software, hardware and a high-capacity information technology infrastructure. There is no general analytical approach that fits all situations; which approach to use depends on the organism and the questions to be answered. We are just beginning to learn what will work in different situations. Interpretation of WGS results must be made

simpler before the method can be implemented routinely in clinical and public health microbiology. Importantly, software must be developed that can be easily used and interpreted by the end-users, e.g., clinical and laboratory personnel with limited insight into genomics. Third, as for any subtyping method, WGS applications need to be validated for reproducibility (repeatability and stability) and discriminatory power for each organism and epidemiologic context under study (5). As these hurdles are overcome the utility of WGS will increase, and it is sure to be widely used in public health and clinical laboratories in the future.

Nonstandard abbreviations

WGS	whole genome sequencing
CDI	*Clostridium difficile* infection
SNV	single nucleotide variant
MLST	multilocus sequence typing

References

1. 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; 369:1195–1205. [PubMed: 24066741]
2. Cohen SH, Gerding DN, Johnson S, Kelly CP, Loo VG, McDonald LC, Pepin J, Wilcox MH. Clinical practice guidelines for *Clostridium difficile* infection in adults: 2010 update by the Society for Healthcare Epidemiology of America (SHEA) and the Infectious Diseases Society of America (IDSA). Infect Control Hosp Epidemiol. 2010; 31:431–55. [PubMed: 20307191]
3. Zhang H, Li D, Zhao L, Fleming J, Lin N, Want T, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013; 45:1255–60. [PubMed: 23995137]
4. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011; 365:709–17. [PubMed: 21793740]
5. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007; S3:1–46.

Table

Advantages and Disadvantages of Three Common Bacterial Subtyping Methods

Subtyping Method	Advantages	Disadvantages
Whole Genome Sequencing (WGS)	Analysis may be tailored for low or high discriminatory power May potentially be automated Provides phylogenetically relevant subtyping data Reproducibility likely to be high if method of data interpretation is standardized, but this requires demonstration in clinical use Typability high Data acquisition methods are applicable to any species without significant adaptation (customizing)	Expensive, but cost is falling Requires technical expertise to perform Labor intensive Requires high informatics capacity and special software Requires bioinformatics expertise, typically doctoral level, to analyze and interpret data; analytical approach must be adapted for each organism and research question Generally long turn-around time, although can be performed in days with concerted effort Currently an experimental method for outbreak investigations
Pulse Field Gel Electrophoresis (PFGE)	Current gold standard for highly discriminatory subtyping, i.e. outbreak investigations Reproducibility high if rigorously standardized Typability high Readily adapted to a variety of species, methods already well described for many species Inexpensive	Fairly labor intensive Requires expertise to interpret the data Do not produce phylogenetically relevant information
Multilocus Sequence Typing (MLST)	Used for phylogenetic subtyping Repeatability, reproducibility high Typability high	Low to moderate discriminatory power, i.e. little use in outbreak investigations Moderately expensive Labor intensive Requires expertise to interpret data Requires significant labor and expertise to adapt and validate for each species