AMERICAN SOCIETY FOR MICROBIOLOGY | Clinical Microbiology Reviews®

REVIEW

# Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis

**Scott Quainoo,[a] Jordy P. M. Coolen,[b] Sacha A. F. T. van Hijum,[c,d] Martijn A. Huynen,[c] Willem J. G. Melchers,[b] Willem van Schaik,[e] Heiman F. L. Wertheim[b]**

Department of Microbiology, Radboud University, Nijmegen, The Netherlands[a]; Department of Medical Microbiology, Radboud University Medical Centre, Nijmegen, The Netherlands[b]; Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands[c]; NIZO, Ede, The Netherlands[d]; Institute of Microbiology and Infection, University of Birmingham, Birmingham, United Kingdom[e]



**Published** 30 August 2017

**Citation** Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. Clin Microbiol Rev 30:1015–1063. https://doi.org/10.1128/CMR.00016-17.



Address correspondence to Scott Quainoo, scott.quainoo@gmail.com.

S.Q. and J.P.M.C. contributed equally to this work.

**SUMMARY** Outbreaks of multidrug-resistant bacteria present a frequent threat to vulnerable patient populations in hospitals around the world. Intensive care unit (ICU) patients are particularly susceptible to nosocomial infections due to indwelling devices such as intravascular catheters, drains, and intratracheal tubes for mechanical ventilation. The increased vulnerability of infected ICU patients demonstrates the importance of effective outbreak management protocols to be in place. Understanding the transmission of pathogens via genotyping methods is an important tool for outbreak management. Recently, whole-genome sequencing (WGS) of pathogens has become more accessible and affordable as a tool for genotyping. Analysis of the entire pathogen genome via WGS could provide unprecedented resolution in discriminating even highly related lineages of bacteria and revolutionize outbreak analysis in hospitals. Nevertheless, clinicians have long been hesitant to implement WGS in outbreak analyses due to the expensive and cumbersome nature of early sequencing platforms. Recent improvements in sequencing technologies and analysis tools have rapidly increased the output and analysis speed as well as reduced the overall costs of WGS. In this review, we assess the feasibility of WGS technologies and bioinformatics analysis tools for nosocomial outbreak analyses and provide a comparison to conventional outbreak analysis workflows. Moreover, we review advantages and limitations of sequencing technologies and analysis tools and present

a real-world example of the implementation of WGS for antimicrobial resistance analysis. We aimed to provide health care professionals with a guide to WGS outbreak analysis that highlights its benefits for hospitals and assists in the transition from conventional to WGS-based outbreak analysis.

**KEYWORDS** bioinformatics, intensive care units, next-generation sequencing, nosocomial infections, outbreak analysis, outbreak management, pathogen surveillance, point of care, whole-genome sequencing

## INTRODUCTION

While several improvements have been made to limit the burden of health care-associated infections, outbreaks of especially-multidrug-resistant (MDR) bacteria still present a frequent threat to vulnerable patient populations in hospitals around the world (1). The EPIC II study, which assessed outcomes and prevalences of infections in 13,796 intensive care unit (ICU) patients worldwide, reported that 36% of ICU patients were infected with MDR bacteria, eventually leading to a doubling of their mortality rate compared to uninfected ICU patients (2). ICU patients are the patient group that is most vulnerable to bacterial infections due to their immune systems being compromised by, for instance, indwelling devices and severe underlying illness. In addition to the vulnerable nature of ICU patients, the prolonged overuse of broad-spectrum antibiotics during and after surgical procedures, inadequate nurse-to-patient ratios, and overcrowding lead to the unintended promotion of MDR bacteria and an eventual increase in the number of bacterial outbreaks in hospitals (3, 4). The increased vulnerability and consequent high mortality rates of infected ICU patients demonstrate the need for effective and standardized outbreak management protocols to be in place (5).

As part of most outbreak management protocols, several phenotypic and molecular methods for pathogen characterization are conventionally used to monitor and curb the spread of resistant bacterial pathogens in hospitals worldwide (6). However, conventional outbreak control approaches often fail to distinguish closely related outbreak strains or detect virulence/resistance features. This is due largely to the limited genomic resolution of conventional molecular methods and the target-specific nature of outbreak analysis approaches; e.g., during infections by antimicrobial-resistant organisms, genotypic tests are employed, which detect only antimicrobial resistance (AMR) genes but not virulence genes, which, if detected concurrently, can provide additional phylogenetic information and improve outbreak analysis (7). To overcome these caveats of conventional outbreak management, novel technologies that provide higher genomic resolution and full genetic information on the entire bacterial genome are needed. Whole-genome sequencing (WGS) can cover all these relevant genomic characteristics, but clinicians have long been hesitant to implement WGS in standard outbreak analysis protocols due to high costs and the cumbersome nature of early next-generation sequencing (NGS) technologies (8–10). Recent advances in sequencing technologies and analysis tools have rapidly increased the output and analysis speed as well as reduced the costs of WGS (11, 12). There is now an ever-increasing body of evidence showing that WGS can provide a fast and affordable outbreak analysis method with a markedly higher resolution than those of conventional methods (13–15). In several countries, such as the United States, Denmark, the United Kingdom, Germany, and The Netherlands, WGS-based pathogen typing is already in the trial phase for implementation as a routine tool for the monitoring and detection of MDR pathogens (16–19) as well as for the early detection of outbreaks (20–22). Still, one has to bear in mind that PCR-based techniques offer relatively cheap and fast typing of isolates and screening for gene functions using dedicated primer sets at a lower resolution.

A number of excellent reviews have covered next-generation sequencing technologies and analysis tools in great detail (23–27). Several important sequencing technologies are not discussed in our review, such as the 454 genome sequencer (Roche) (8),

the Ion Torrent personal genome machine (Life Technologies) (9), and the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system (Applied Biosystems) (10), as they have been superseded by other sequencing technologies. Instead, we assessed the performances of today's most frequently used sequencing technologies as well as the latest developments in sequencing technologies. Furthermore, the performances of selected bioinformatics tools for assembly, genome characterization, comparative genomics, and phylogeny were reviewed. In an attempt to provide a representative overview of the vast number of bioinformatics tools to a broad audience, our analysis included both well-established and recently developed algorithms, which span over three different user interface types and require various levels of bioinformatics skills. Finally, we discuss the benefits and drawbacks of using the selected sequencing technologies and analysis tools and provide a future outlook for the real-world implementation of WGS-based outbreak analyses.

## OUTBREAK DEFINITION

According to the Centers for Disease Control and Prevention (CDC), an outbreak is defined as "the occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time" (https://www.cdc.gov/). Instead of disease, one may also consider the state of carrying a specific pathogen, such as a multidrug-resistant *Pseudomonas* strain. An outbreak alert might be triggered by a cluster of patients colonized with the same drug-resistant Gram-negative bacterium (GNB) in an ICU ward (3). According to a study by Gastmeier et al., which reviewed the 2005 worldwide database of health care-associated outbreaks (https://www.outbreak-database.com/), outbreaks in neonatal ICUs are due mainly to *Klebsiella* spp. (20.3%) and *Staphylococcus* spp. (15.9%), with the majority of infections being bloodstream infections (62.7%) and gastrointestinal infections (20.7%) (28). In other ICUs, the majority of infections are due largely to *Staphylococcus* spp. (20.1%) and *Acinetobacter* spp. (15.9%), with the majority of infections being bloodstream infections (46.8%) and pneumonia (20.7%) (28). The majority of infection sources are reportedly unknown, followed by infections originating from patients, the environment, medical equipment, and health care personnel (28, 29).

## CONVENTIONAL MOLECULAR CHARACTERIZATION METHODS

For many years, the large majority of clinical microbiology laboratories used several methods for characterizing bacterial strains, including serotyping (30, 31), antimicrobial susceptibility testing (32, 33), and mass spectrometry (MS)-based (34) methods that are still considered the gold standard of phenotypic characterization of pathogenic bacteria. In an extensive review, van Belkum et al. provide a detailed description of conventional phenotypic and molecular characterization methods (6). While conventional phenotypic characterization methods have proven to be successful in identifying and controlling outbreaks in ICUs, they all have the common disadvantage of being time-consuming and providing low taxonomic resolution (35, 36). In recent years, pathogen characterization has therefore moved to more sensitive genomic analysis techniques. The early beginnings of genomic analysis were made by the use of several genetic analysis tools that focus on small parts of the bacterial genome (6). In the focus of our review, the most frequently used non-amplification- and amplification-based genomic methods are described briefly.

### Non-Amplification-Based Typing Technologies

**Restriction fragment length polymorphism methods.** In restriction endonuclease analysis (REA), one of the first restriction fragment length polymorphism (RFLP) methods, a bacterial chromosome is subjected to a digestion step, where restriction enzymes cut the chromosome into smaller fragments, which are then separated by size via gel electrophoresis (37). Under a standardized protocol, this method is relatively fast, discriminatory, and easy to reproduce, yet the complex nature of the produced patterns makes interpretation of the results difficult and hampers data exchange between

different research groups (6). To improve the interpretation of results, a combination of RFLP and ribotyping can be used, where, in addition to genome digestion, a second step is added, which hybridizes an rRNA gene-complementary probe to the genome fragments. Certain hybridization probes that are species specific can be used, such as during IS*6110* typing, in which standardized typing of *Mycobacterium tuberculosis* can be achieved (38). However, despite these improvements, studies have shown that RFLP clusters lack discriminatory power and can be further subdivided by newer WGS-based typing methods (39, 40). The higher resolution of such WGS methods could enable clinicians to better distinguish outbreak strains from nonoutbreak strains.

Several other non-amplification-based methods are commonly used, such as DNA-DNA reassociation, which assesses the hybridization of DNA fragment pools to infer genetic distances between organisms (41), and plasmid typing, which distinguishes bacteria based on their unique profiles of plasmids (42).

**Matrix-assisted laser desorption ionization–time of flight mass spectrometry.** Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) MS (43) is a molecular typing technique that identifies bacterial isolates based on unique protein profiles. For detection, a protein spectrum is obtained and compared to a reference database of bacterial protein spectra to identify the isolate. MALDI-TOF MS has been established as a frequently used method for the identification of bacterial pathogens during routine screenings (44, 45) and for the distinction of bacterial strains during nosocomial outbreaks in intensive care units (46–48). For an extensive description of further applications of MALDI-TOF MS in microbiological diagnostics, the reader is referred to a review by Wieser et al. (49).

Recently, Schlebusch et al. described the complementary use of MALDI-TOF MS and WGS for the investigation of a vancomycin-resistant *Enterococcus faecium* (VRE) outbreak (50). That study highlighted the inconsistency of MALDI-TOF MS results based on potential biases in phenotypic typing data from various protein expression levels. Even though MALDI-TOF MS was able to distinguish outbreak strains with shorter turnaround times (TATs), WGS analysis provided far-higher discriminatory power, which ultimately allowed an improved understanding of transmission events. That study hence argued that in an outbreak scenario, MALDI-TOF MS could be used to complement WGS as a rapid initial analysis tool until WGS data are generated.

**Pulsed-field gel electrophoresis.** Pulsed-field gel electrophoresis (PFGE) is a typing technique that differentiates bacterial isolates at the strain level. During PFGE, a fingerprint (pulsotype) of DNA fragments is generated on a gel and compared to a database, the extent of which can vary largely depending on the bacterial species, to identify the bacterial isolate (51).

A major disadvantage of this method is the inconsistency in results caused by the use of multiple standardized protocols and variations of restriction enzymes from the same or between different manufacturers (52). However, PFGE networks such as PulseNET present examples where the coordinated implementation of standardized workflows can result in the successful implementation of this technique at the national level (53).

Despite its widely accepted use as a highly sensitive typing method, PFGE is a laborious method due to its finicky sample preparation, long run time, and dependence on bacterial culture (51, 54). Even though the costs of PFGE are still approximately half of those associated with newer WGS-based typing methods (55), the superiority of WGS over PFGE in bacterial typing has been successfully demonstrated in analyses of bacterial transmission events. Several studies have shown the higher discriminatory power of WGS than of PFGE in identifying transmission events during outbreaks of methicillin-resistant *Staphylococcus aureus* (56) and *Escherichia coli* O157:H7 (57) infections.

## Amplification-Based Typing Technologies

**Multiple-locus variable-number tandem-repeat analysis.** The limitations of PFGE have led to the development of cheaper, faster, and more detailed PCR-based typing

methods. Multiple-locus variable-number tandem-repeat (VNTR) analysis (MLVA) is a typing method that discriminates closely related bacterial strains based on their numbers of VNTRs. PCR primers are chosen to be outside the VNTR region, producing DNA fragments of various lengths depending on the number of repeats. PCR products are then analyzed through capillary electrophoresis to determine their size via the use of specific software. Results are usually reported as a string of numbers representing the VNTRs at each tested locus (58), allowing universal interpretation. One example of MLVA is *spa* typing, where strains of *S. aureus* are discriminated based on the staphylococcal protein A (*spa*) gene (59).

PCR-based MLVA was demonstrated to be a faster and more available alternative to PFGE, as it is able to discriminate between highly related bacterial strains. However, Bertrand et al. demonstrated that for clinical isolates of *Salmonella enterica* serovar Enteritidis, it was possible with other typing methods to further discriminate the most common MLVA profile identified into five phage subtypes (58). Hence, when investigations are performed on bacterial isolates with a highly common MLVA profile, the technique should be accompanied by complementary typing methods, such as WGS-based approaches, to achieve unique subtyping results and increased resolution. In fact, WGS-based typing has been shown to be less expensive, less labor-intensive, and of higher resolution for strain distinction than MLVA (60).

**Multilocus sequence typing.** Multilocus sequence typing (MLST) is a typing technique that identifies bacteria based on sequence differences in housekeeping genes. MLST can be performed through either a single-gene sequencing or a WGS approach; a detailed description of the latter follows later in this review. For usually at least seven housekeeping genes, the sequence differences for a bacterial isolate are assigned a distinct allele. The alleles at each of the loci (genes) are described as the allelic profile or sequence type (ST). This ST can then be used as a barcode to differentiate isolates and establish evolutionary relationships via designated analysis tools (61).

MLST has been shown to be effective in identifying pathogenic bacterial strains with high resolution (62, 63); however, the high level of variation of housekeeping genes among different bacteria makes it possible to create MLST schemes only for bacterial pathogens that are highly related at the genus-to-species levels (61). MLST furthermore does not provide discrimination between variants of a single clone, which is relevant for asexual pathogens such as *Bacillus anthracis* and *Yersinia pestis*, which can make this method insufficient as an outbreak analysis tool for such pathogens (64). In organisms with considerable levels of recombination, the same MLST type may hide considerable biological diversity, which may result in inappropriate conclusions on the clonal nature of strains (65–67).

**Virulence gene typing.** In addition to typing, PCR can be used to identify bacterial pathogens based on specific virulence factors such as toxins, adhesins, or capsules. As in PCR-based genotyping, species-specific virulence genes are assessed as PCR primer targets and amplified for the characterization of a pathogen in a sample (68–70). Traditional PCR detection of virulence genes has the disadvantage of being able to identify only one gene or species per reaction, which limits its use in high-throughput outbreak analyses. Multiplex PCR methods have hence been established to detect multiple species and genes in one sample with the use of multiple target-specific primers. The multiplex method is a well-established method for the fast and reliable detection of virulence genes and has been shown by several studies to be successful in detecting virulence, antibiotic resistance, and toxin (VAT) genes in *Campylobacter* species and virulence-associated genes in *Arcobacter* species, to name only a few examples (69, 71). However, limitations in resolution and the superiority of WGS over PCR-based detection of virulence genes at comparable TATs have been demonstrated (20). Therefore, WGS-based detection of virulence genes might be more suitable than PCR-based methods in outbreak situations where high-resolution detection of virulence determinants could lead to improved pathogenicity characterization and, consequently, outbreak control.

In addition to the methods described above, several other amplification-based

methods are used for pathogen characterization, such as amplified rRNA restriction analysis, a modified RFLP method that analyzes the 16S rRNA gene (72); random amplified polymorphic DNA (RAPD) analysis, where PCR using arbitrary primers amplifies random DNA sequences to create a semiunique DNA fragment profile for isolate identification (73); and amplified fragment length polymorphism (AFLP), a PCR method that amplifies restriction fragments from genomic DNA digests to create DNA fingerprints for the identification of bacterial isolates (74).

## NEED FOR WGS FOR OUTBREAK ANALYSIS

The above-described amplification-based and non-amplification-based methods are used to investigate only small fragments of the bacterial genome, which limits these approaches to species-dependent protocols. WGS-based typing of bacterial pathogens includes mobile genetic elements and could provide unprecedented resolution in discriminating even highly related lineages, thereby obviating the use of species-dependent protocols. By sequencing the entire genome (chromosome and mobile genetic elements), WGS immediately provides information on pathogen detection and identification, epidemiological typing, and drug susceptibility, which is crucially important information that in conventional outbreak management is achievable only through the use of multiple methods.

Of additional importance is the fact that resistance/virulence genes detected via WGS might not be expressed under conditions of phenotypic testing *in vitro* or, for that matter, *in vivo*. In particular, there have been reports of the "*in vivo*-only" expression of virulence gene promoters in *S. aureus* and *Salmonella enterica* serovar Typhimurium (75, 76). The detection of such pathogenicity features via WGS could help clinicians identify potential nosocomial transmission events earlier and manage bacterial outbreaks before conventional phenotypic tests can detect them.

Despite the concerns of high operational costs associated with WGS, which are frequently voiced by health care professionals (77–79), WGS pipelines could potentially reduce overall costs for hospital practices through savings of indirect costs. Of note is a recent study by Mellmann et al., which assessed the performance of a novel WGS typing pipeline for monitoring bacterial transmission in a multibed-room, tertiary hospital in Germany (55). That study successfully demonstrated that WGS typing was more precise in excluding the majority of bacterial isolates from nosocomial transmission clusters than conventional typing methods such as PFGE. These results prompted a reduction in the number of patient isolation procedures over a 6-month period, which in turn enabled cost savings of more than $230,000, largely due to reduced workloads and indirect savings from the avoidance of blocked beds.

## METHODS

For this review, sequencing technologies were assessed based on sequence coverage, output quantity, consumables and instrument costs, read length, number of reads per run, cost per gigabase, run time, and error rates. Sequencing coverage describes the average number of aligned read fragments that cover a specific nucleotide in the reconstructed sequence and is calculated by dividing the total output by the target genome size and dividing this result by the number of samples per run. To provide examples of coverage for each sequencing technology, this review calculated coverage based on the genome size of *S. aureus* strain MRSA252. Presented coverages can then be compared to reference values of 35-fold to 50-fold for small genomes, as previously recommended (80). Output describes the amount of sequence information produced per sequencing run. Error rates were analyzed from reported benchmarks of "raw" sequence data after a sequencing run was completed. As possible improvements in error values through data cleaning can vary highly depending on data sets, sequencing technology, and sample preparation, etc., we decided not to mention error values after additional improvement of the data. By doing so, this review aims to present the reader with an unbiased picture of the machine performance of each technology described.

Tools for the analysis of WGS data were divided into five groups: assembly, genome

characterization, comparative genomics, phylogeny, and complete outbreak analysis software suites. Assembly tools were assessed based on sequencing technology, computational requirements, speed, and assembly quality. Computational requirements were based on the reported random-access memory (RAM) usage for various benchmarking data sets, speed was based on the reported run time for various benchmarking data sets, and assembly quality was based on reported $N_{50}$ values and percentages of identity for various benchmarking data sets. In a given set of assembled contigs, the $N_{50}$ value describes the base pair length of the shortest contig in an assembly, such that the sum of all contigs of longer or identical lengths results in a minimum of half the total base pair length of all contigs of the original assembly. Genome characterization tools were assessed mainly based on input/output types. Tools for comparative genomics and phylogeny estimations were assessed based on input/output type, run time, and topology score/accuracy. The complete outbreak analysis software suites were assessed based on RAM compatibility, the number of schemes, price, and run time.

## SEQUENCING TECHNOLOGIES

Ever since the first report of a complete bacterial genome sequence in 1995 (81), sequencing technologies have rapidly improved. As presented in Table 1, second-generation sequencing platforms allow whole bacterial genomes to be sequenced within hours, while third-generation sequencing platforms, that provide longer reads and additional information, such as methylation sites, with even higher speed have been developed (82). This review assesses the performance of popular sequencing platforms as well as emerging state-of-the art technologies that were available at the time of writing of this review. The results of the performance assessment are shown in Table 1.

### Illumina

**Principle of technology.** The Illumina sequencing platforms use fluorescently labeled nucleotides (deoxynucleoside triphosphates [dNTPs]) to determine the genetic sequence of DNA fragments. Here we focus on three Illumina model series: MiniSeq, the smallest, most affordable Illumina sequencer; MiSeq, a simple system for rapid sequencing with relatively low outputs; and NextSeq, a midsized, flexible system with options for high- and mid-range outputs.

The Illumina sequencing-by-synthesis (SBS) technology begins with several library preparation steps (83). Initially, purified sample DNA is fragmented by either mechanical shearing, e.g., via sonication, or enzymatic shearing, e.g., via transposases. Unique adaptor sequences (and, optionally, barcodes) are then ligated to either end of the DNA fragments and loaded onto a reagent cartridge that is inserted into the sequencer. The sequencer then loads the mix of reagents and DNA fragments into a solid-surface flow cell that is coated with primers complementary to the adaptor sequences. The ligated fragment ends then bind to the cell surface, and a DNA polymerase amplifies the fragments to produce several copies of the initial DNA fragment, called clusters. Next, four different fluorescently labeled nucleotides (A, C, G, and T) are added to the flow cell and incorporated by a polymerase into a new DNA strand one base at a time. The MiniSeq and NextSeq systems use a two-fluorophore system, instead of the four-fluorophore system used by the MiSeq system (23). After a wash step, the fluorescence of incorporated nucleotides is imaged by using one of four different imaging channels. Next, the fluorescent dyes are cleaved off and washed away, and the process is repeated. The sequencer documents the color changes after nucleotide addition to construct the genetic sequence of the DNA clusters. Either results can be analyzed as single-end reads or a second strand can be synthesized, and the process is repeated for paired-end reads. Paired-end reads provide more sequencing information but increase the sequencing cost and time needed for sequencing.

**Specifications.** Whereas enzymatic reactions take very little time, the major contributor to run time is the imaging of the flow cell. Illumina has reduced the run time of previous models considerably by reducing the imaged surface area on the flow cell. As

**TABLE 1** Performance analysis of sequencing platforms[a]

| Platform | Read length (bp) | Output (Gb) | Coverage[b] | Run time (h) | No. of reads | Cost per Gb ($) | Consumables cost ($) | Instrument cost ($) | Error rate | Dimensions (width × depth × ht) (cm) | Source(s) (reference[s]) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequencing by synthesis | | | | | | | | | | | |
| Illumina MiniSeq Mid Output | 2 × 150[c] | 2.1–2.4[c] | 8.6 | 17[c] | 14 million–16 million[c] | 2,584–2,953[d] | 6,201[c] | 55,411[c] | 0.1% in >80% of base calls[c] | 45.6 × 48 × 51.8[c] | Illumina |
| Illumina MiniSeq High Output | 1 × 75[c] | 1.7–1.9[c] | 6.8 | 7[c] | 22 million–25 million[c] | 3,264–3,648[d] | 6,201[c] | 55,411[c] | 0.1% in >85% of base calls[c] | 45.6 × 48 × 51.8[c] | Illumina |
| | 2 × 75[c] | 3.3–3.8[c] | 13.6 | 13[c] | 44 million–50 million[c] | 1,632–1,879[d] | 6,201[c] | 55,411[c] | 0.1% in >85% of base calls[c] | 45.6 × 48 × 51.8[c] | Illumina |
| | 2 × 150[c] | 6.6–7.5[c] | 26.9 | 24[c] | 44 million–50 million[c] | 827–940[d] | 6,201[c] | 55,411[c] | 0.1% in >80% of base calls[c] | 45.6 × 48 × 51.8[c] | Illumina |
| Illumina MiSeq Reagent kit v2 | 1 × 36[c] | 0.54–0.61[c] | 2.2 | 4[c] | 12 million–15 million[c] | 7,946–8,976[d] | 4,847[c] | 108,244[c] | 0.1% in >90% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| | 2 × 25[c] | 0.75–0.85[c] | 3.1 | 5.5[c] | 24 million–30 million[c] | 5,702–6,463[d] | 4,847[c] | 108,244[c] | 0.1% in >90% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| | 2 × 150[c] | 4.5–5.1[c] | 18.3 | 24[c] | 24 million–30 million[c] | 950–1,077[d] | 4,847[c] | 108,244[c] | 0.1% in >80% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| | 2 × 250[c] | 7.5–8.5[c] | 30.5 | 39[c] | 24 million–30 million[c] | 570–646[d] | 4,847[c] | 108,244[c] | 0.1% in >75% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| Illumina MiSeq Reagent kit v3 | 2 × 75[c] | 3.3–3.8[c] | 13.6 | 21[c] | 44 million–50 million[c] | 1,362–1,568[d] | 5,174[c] | 108,244[c] | 0.1% in >85% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| | 2 × 300[c] | 13.2–15[c] | 53.8 | 56[c] | 44 million–50 million[c] | 345–392[d] | 5,174[c] | 108,244[c] | 0.1% in >70% of base calls[c] | 68.6 × 56.5 × 52.3[c] | Illumina |
| Illumina NextSeq 500 Mid Output | 2 × 75[c] | 16.3–20[c] | 71.8 | 15[c] | <260 million[c] | 318–391[d] | 6,369[c] | 266,835[c] | 0.1% in >75% of base calls[c] | 53.3 × 63.5 × 58.4[c] | Illumina |
| | 2 × 150[c] | 32.5–39[c] | 140 | 26[c] | <260 million[c] | 163–196[d] | 6,369[c] | 266,835[c] | 0.1% in >80% of base calls[c] | 53.3 × 63.5 × 58.4[c] | Illumina |
| Illumina NextSeq 500 High Output | 1 × 75[c] | 25–30[c] | 107.7 | 11[c] | <400 million[c] | 312–374[d] | 9,347[c] | 266,835[c] | 0.1% in >80% of base calls[c] | 53.3 × 63.5 × 58.4[c] | Illumina |
| | 2 × 75[c] | 50–60[c] | 215.3 | 18[c] | <800 million[c] | 156–187[d] | 9,347[c] | 266,835[c] | 0.1% in >80% of base calls[c] | 53.3 × 63.5 × 58.4[c] | Illumina |
| | 2 × 150[c] | 100–120[c] | 430.6 | 29[c] | <800 million[c] | 78–93[d] | 9,347[c] | 266,835[c] | 0.1% in >75% of base calls[c] | 53.3 × 63.5 × 58.4[c] | Illumina |
| Single-molecule real-time sequencing | | | | | | | | | | | |
| Pacific Biosciences RS II P6-C4 chemistry | >20,000[c] | 8–16[e] | 57.4 | 0.5–4[c] | 55,000[c] | 250–500[d] | 4,000[e] | 695,000 | 14% errors per base | 203.0 × 90.0 × 160.0[c] | PacBio, AllSeq[f] (89) |
| Pacific Biosciences Sequel system | >20,000[c] | 80–160[e] | 574.2 | 0.5–6[c] | 370,000[c] | 70–140[d] | 11,200[e] | 350,000 | 14% errors per base | 92.7 × 86.4 × 167.6[c] | PacBio, AllSeq[f] (89) |
| Oxford Nanopore MinION Mk1 (1D) | >882,000 | 10–20[c] | 71.8 | 1.67–72[c] | 138,000 | 49.95–99.9[d] | 999[c] | 1,000[c] | 12% errors per base | 10.5 × 3.3 × 2.3[c] | Oxford Nanopore Technologies, Loman Labs[g] (231) |
| Oxford Nanopore MinION Mk1 (2D) | >882,000 | 10–20[c] | 71.8 | 1.67–72[c] | 138,000 | 49.95–99.9[d] | 999[c] | 1,000[c] | 15% errors per base | 10.5 × 3.3 × 2.3[c] | Oxford Nanopore Technologies, Loman Labs[g] (231, 232) |
| Oxford Nanopore PromethION single flow cell | <300,000[c] | 233[c] | 836.2 | 1.67–>72[c] | 26 million[c] | NA | NA | 135,000 (PEAP)[c] | NA | 44.0 × 24.0 × 40.0[c] | Oxford Nanopore Technologies |
| Oxford Nanopore PromethION 48 flow cells | <300,000[c] | 11,000[c] | 3,9475.8 | 1.67–>72[c] | 1.25 billion[c] | NA | NA | 135,000 (PEAP)[c] | NA | 44.0 × 24.0 × 40.0[c] | Oxford Nanopore Technologies |

[a]All quantitative performance measures were taken from previously reported data, as indicated. Consumables costs were calculated as follows: Illumina costs included PhiX Control kit v3, the Nextera XT DNA sample preparation kit (96 samples)/Nextera DNA library preparation kit (96 samples), and Nextera XT Index kit v2 (96 indexes and 384 samples), the highest-output reagent kit. PEAP, PromethION Early-Access Program; NA, no data available.

[b]Calculated for 96 samples and the genome size of *S. aureus* strain MRSA252 (2,902,619 bp).

[c]Manufacturer's data.

[d]Estimated calculation for consumables.

[e]For 16 SMRT cells.

[f]See http://www.allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/.

[g]See http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/.

shown in Table 1, total run times, including cluster generation, sequencing, and base calling, can hence be reduced on the Illumina MiSeq system to 4 h and 56 h at the lowest-output (reagent kit v2) and highest-output (reagent kit v3) settings, respectively. However, with a decrease in the imaged surface area, the total number of generated data points per run decreases, which in turn increases the sequencing cost per nucleotide considerably (24).

On the fastest setting, the MiSeq system (reagent kit v2) can produce a minimum of 0.54 to 0.61 Gb of data with a single-end read length of 36 bp. On the more powerful NextSeq 500 system, a data output of 100 to 120 Gb can be achieved in the highest-output mode with a paired-end read length of 150 bp.

The average sequencing cost presented here is either taken from the literature or estimated based on the listed prices for consumables and output by the manufacturer, as indicated in Table 1. Most Illumina sequencing machines require a PhiX DNA control kit, a DNA library preparation kit, an indexing primer kit to allow the sequencing of up to 96 pooled samples, and a reagent kit. The sequencing costs per gigabase decrease with higher total outputs and hence start from $7,946 to $8,976/Gb with the MiSeq system (reagent kit v2, 1- by 36-bp read length) and can be decreased to around $78 to $93/Gb with the NextSeq 500 system (high output, 2- by 150-bp read length), the latter of which is the lowest range of sequencing costs per gigabase of the sequencers described in this study. Here it must be noted that multiple bacterial genomes can be run on the Illumina sequencers at a time, which reduces the costs per genome accordingly. As shown in Table 1, Illumina sequencers are offered at competitive instrument prices compared to those of other technologies, such as those of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). With prices ranging from $55,411 to $266,835 for the Illumina MiniSeq and the Illumina NextSeq 500 systems, respectively, instrument costs are lower than those of the PacBio system but well above those of the cheapest ONT sequencers. The relation between instrument cost and other parameters, such as instrument footprint, is an important aspect to consider when evaluating the costs of WGS infrastructures for specific hospital needs.

On Illumina systems, error rates in base calling are predicted by a quality score. A quality score of 30 ($Q_{30}$) predicts an error rate of 0.1 or an error of 1 in 1,000 base callings. The MiSeq system (reagent kit v2) achieves the highest quality score, 0.1% for >90% of base callings, and the MiSeq system (reagent kit v3) produces the lowest score, 0.1% for >70% of base callings.

The Illumina platforms have already been used for pathogen detection during outbreaks, and several studies have demonstrated their applicability and superiority over conventional methods in terms of outbreak control in clinical settings. A study by McGann et al. used WGS to study an outbreak of VRE that occurred among three ICU patients at a tertiary care hospital in Honolulu, HI (84). TATs for the Illumina MiSeq sequencer were determined to assess its applicability in a clinical setting during outbreaks. The initial epidemiological assessment was based on the timeline of the outbreak and suggested linear nosocomial transmission of the outbreak pathogen from a source patient (patient A) to a second patient (patient B) and, consequently, to a third patient (patient C). However, in contrast to the initial assessment, sequence data generated on the Illumina MiSeq system revealed that isolates of patient A differed from the isolates of the two other patients (patients B and C) by one single nucleotide polymorphism (SNP). This indicated that instead of the initially suspected linear transmission route, two separate events of transmission from patient A to both patients B and C most likely occurred. WGS therefore improved the understanding of the outbreak transmission network, which, in retrospect, could have potentially enhanced the outbreak control response at that time. The sequencer provided superior resolution with a TAT, including overnight culturing, of 48.5 h, which would allow a faster and more comprehensive response by infection control teams than with conventional detection methods with TATs of several weeks (15).

Another evaluation of the use of WGS for outbreak surveillance was recently conducted by Kwong et al. in the context of *Listeria monocytogenes* surveillance in

Australia (60). That study compared the performance of WGS via the Illumina NextSeq or MiSeq system to those of conventional typing methods, including binary typing, PCR serotyping, MLST, MLVA, and PFGE. Besides being highly concordant (>99%) with results of binary typing, MLST, and serotyping, WGS enabled the identification of separate nested clusters among isolate groups that were undetectable with conventional methods. During additional routine epidemiological surveillance over a 12-month period, WGS allowed higher resolution in linking point source outbreaks than conventional typing. Based on these results, Kwong and colleagues were able to develop a nationwide risk-based alert system for WGS data to inform epidemiologists of sequence similarities and possible events of transmission of bacterial pathogens at discriminatory powers far superior to those with conventional typing-based surveillance.

### Pacific Biosciences

**Principle of technology.** While Illumina sequencers have proven their accurate performance, there are limitations in their short reads, creating problems with the determination and assembly of complex genomic regions. PacBio's third-generation sequencing platforms, the Sequel system and RSII, aim to solve this issue by implementing single-molecule real-time (SMRT) sequencing (85). The SMRT technology achieves this in two main steps. First, a so-called SMRT bell is generated by ligating both ends of a double-stranded target DNA with hairpin adaptors. The SMRT bell is then loaded onto a SMRT cell that contains a number of microscopic chambers, called zero-mode wave guides (ZMWs), that act as a detection space during sequencing. As the SMRT bell is loaded onto the cell, its hairpin adaptor binds to an immobilized DNA polymerase at the bottom of the ZMW. Next, fluorescently labeled nucleotides (A, C, G, and T) are added to the cell. As the polymerase begins to incorporate labeled nucleotides into a new DNA strand, the fluorescent labels are cleaved off and produce light pulses of emission spectra unique to each base. The light pulses are detected by a laser beam and recorded in real time to determine the nucleotide sequence as a continuous long read (CLR) (86). With this technology, it is possible to simultaneously detect thousands of single-molecule sequencing reactions at high speeds. Whereas the individual light signals are recorded in real time, the data cannot be observed in real time unless the run is stopped for observation.

**Specifications.** Due to the lack of amplification, SMRT sequencing makes the PacBio sequencers some of the fastest on the market, with total run times of 0.5 to 4 h on the RSII (P6-C4) system. This makes the technology extremely valuable for outbreak analyses, where quick identification leads to faster treatment and, eventually, reductions in costs and loss of life.

As shown in Table 1, the output of PacBio systems is one of the lowest available on the market, with only 500 Mb to 1 Gb per SMRT cell on the RSII (P6-C4) system and 5 to 10 Gb per SMRT cell on the Sequel system. However, as indicated in Table 1, both the RSII and Sequel systems allow the running of up to 16 SMRT cells at once, which increases total outputs. The low output is due mainly to the focus on long reads for genome assembly, making it possible to achieve read lengths of >20 kb.

The sequencing costs per gigabase for PacBio sequencers are comparatively cheap, at $250 to $500 for the RSII (P6-C4) system and $70 to $140 with the Sequel system. However, the sequencers are expensive, at $695,000 for the RSII (P6-C4) system and $350,000 for the Sequel system, making PacBio technology one of the costlier options for clinical outbreak analysis.

One error specific to this technique is that during DNA replication in the ZMW, detection of nucleotides that are dwelling long enough at the active site of the polymerase can occur without these nucleotides actually being incorporated into the new DNA strand. These errors accumulate during the sequencing run and increase the overall error rate of the final read (87). Whereas the SMRT sequencing technique allows some of the longest reads available today, the small number of reads per run