



# Validation and Implementation of Clinical Laboratory Improvements Act-Compliant Whole-Genome Sequencing in the Public Health Microbiology Laboratory

Varvara K. Kozyreva, Chau-Linda Truong, Alexander L. Greninger, John Crandall, Rituparna Mukhopadhyay, Vishnu Chaturvedi

Microbial Diseases Laboratory, California Department of Public Health, Richmond, California, USA

**ABSTRACT** Public health microbiology laboratories (PHLs) are on the cusp of unprecedented improvements in pathogen identification, antibiotic resistance detection, and outbreak investigation by using whole-genome sequencing (WGS). However, considerable challenges remain due to the lack of common standards. Here, we describe the validation of WGS on the Illumina platform for routine use in PHLs according to Clinical Laboratory Improvements Act (CLIA) guidelines for laboratory-developed tests (LDTs). We developed a validation panel comprising 10 *Enterobacteriaceae* isolates, 5 Gram-positive cocci, 5 Gram-negative nonfermenting species, 9 *Mycobacterium tuberculosis* isolates, and 5 miscellaneous bacteria. The genome coverage range was 15.71× to 216.4× (average, 79.72×; median, 71.55×); the limit of detection (LOD) for single nucleotide polymorphisms (SNPs) was 60×. The accuracy, reproducibility, and repeatability of base calling were >99.9%. The accuracy of phylogenetic analysis was 100%. The specificity and sensitivity inferred from multilocus sequence typing (MLST) and genome-wide SNP-based phylogenetic assays were 100%. The following objectives were accomplished: (i) the establishment of the performance specifications for WGS applications in PHLs according to CLIA guidelines, (ii) the development of quality assurance and quality control measures, (iii) the development of a reporting format for end users with or without WGS expertise, (iv) the availability of a validation set of microorganisms, and (v) the creation of a modular template for the validation of WGS processes in PHLs. The validation panel, sequencing analytics, and raw sequences could facilitate multilaboratory comparisons of WGS data. Additionally, the WGS performance specifications and modular template are adaptable for the validation of other platforms and reagent kits.

**KEYWORDS** bacteria, whole-genome sequencing, performance specifications, laboratory-developed test, quality management, validation, CLIA, public health, bioinformatics pipeline, WGS

Clinical microbiology laboratories and public health microbiology laboratories (PHLs) are undergoing transformative changes with the adoption of whole-genome sequencing (WGS) (1, 2). For several years, leading laboratories have reported proof-of-concept studies on WGS-enabled advances in the identification of pathogens, antibiotic resistance (ABR) detection, and disease outbreak investigations (3–6). Technologies also referred to as next-generation sequencing (NGS) have yielded more detailed information about the microbial features than was possible by using a combination of other laboratory approaches. Further developments of WGS platforms allowed remarkable in-depth inquiry of pathogenic genomes for the discovery of genetic variants and

Received 2 March 2017 Returned for  
modification 25 March 2017 Accepted 26  
May 2017

Accepted manuscript posted online 7 June  
2017

**Citation** Kozyreva VK, Truong C-L, Greninger AL, Crandall J, Mukhopadhyay R, Chaturvedi V. 2017. Validation and implementation of Clinical Laboratory Improvements Act-compliant whole-genome sequencing in the public health microbiology laboratory. *J Clin Microbiol* 55:2502–2520. <https://doi.org/10.1128/JCM.00361-17>.

**Editor** Daniel J. Diekema, University of Iowa College of Medicine

**Copyright** © 2017 Kozyreva et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Vishnu Chaturvedi, Vishnu.Chaturvedi@cdph.ca.gov.

genome rearrangements that could have been missed by using other DNA methods (3, 7, 8). Enhanced investigations of disease outbreaks have led to a new understanding of routes of transmission of infectious agents (9–11). WGS-enabled metagenomics and microbiome discoveries have revealed a new appreciation for the role of microbes in health and disease (12–15). The innovations are continuing at such an unprecedented pace that WGS is expected to become an alternative to culture-dependent approaches in clinical and public health microbiology laboratories (16–18).

Notwithstanding its promises, several challenges remain for the adoption of WGS in microbiology laboratories (19–22). The accelerated obsolescence of sequencing platforms presents several obstacles in bridging the gap between research and routine diagnostics, including standardization efforts (23). The downstream bioinformatics pipelines are also unique challenges for microbiology laboratories regarding both infrastructure and skilled operators (24–27). Overall, WGS “wet-bench” and “dry-bench” workflows represent integrated processes, which are not easily amenable to the traditional quality metrics used by microbiology laboratories (27–29). The capital investments and recurring costs of WGS for clinical laboratories, although rapidly declining, remain relatively high to allow multilaboratory comparisons for the standardization of analytical parameters. Finally, regulatory agencies have not yet proposed standard WGS guidelines for clinical microbiology (30), and external proficiency testing (PT) programs for clinical and public health microbiology laboratories are still in development (31, 32).

There are a few notable developments toward the standardization and validation of next-generation sequencing in clinical laboratories. The U.S. Centers for Disease Control and Prevention (CDC) sponsored the Next-Generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) workgroup to propose quality laboratory practices for the detection of DNA sequence variations associated with heritable human disorders (33, 34). This workgroup developed principles and guidelines for test validation, quality control, proficiency testing, and reference materials. Although not focused on infectious diseases, these guidelines provide a valuable roadmap for the implementation of WGS in clinical microbiology and public health laboratories. The College of American Pathologists (CAP) reported 18 requirements in an accreditation checklist for next-generation sequencing analytic (wet-bench) and bioinformatics (dry-bench) processes as part of its molecular pathology checklist (30). These “foundational” accreditation requirements were designed to be broadly applicable to the testing of inheritable disorders, molecular oncology, and infectious diseases. Along the same lines, the feasibility of *in silico* proficiency testing has been demonstrated for NGS (35). Recently, high accuracy and reproducibility were shown for WGS-based microbial strain typing performed in a ring trial study involving five laboratories; the results suggested that a proficiency testing program for WGS is feasible in clinical microbiology laboratories (36). The Clinical and Laboratory Standards Institute (CLSI) updated its guidelines for nucleic acid sequencing methods in diagnostic laboratory medicine with considerations specific to the application of next-generation sequencing in microbiology (37). Thus, a broad technical framework is now available to design WGS validation protocols that will be most relevant for clinical and public health laboratories. Our aims for the present study were to establish performance metrics for the typical workflow in public health microbiology laboratories, design modular templates for the validation of different platforms and chemistries, finalize a user-friendly report format, and identify a set of bacterial pathogens that could be used for WGS validation and performance assessments.

## RESULTS

**Accuracy of WGS.** A number of Clinical Laboratory Improvements Act (CLIA)-required performance parameters were adopted, with modifications for the validation of WGS (Table 1). The modular validation template and a summary are presented in Fig. 1. The quality assurance (QA) and quality control (QC) measures are described in detail later in Results.

**TABLE 1** Performance characteristics, definitions, and formulas used for validation<sup>a</sup>

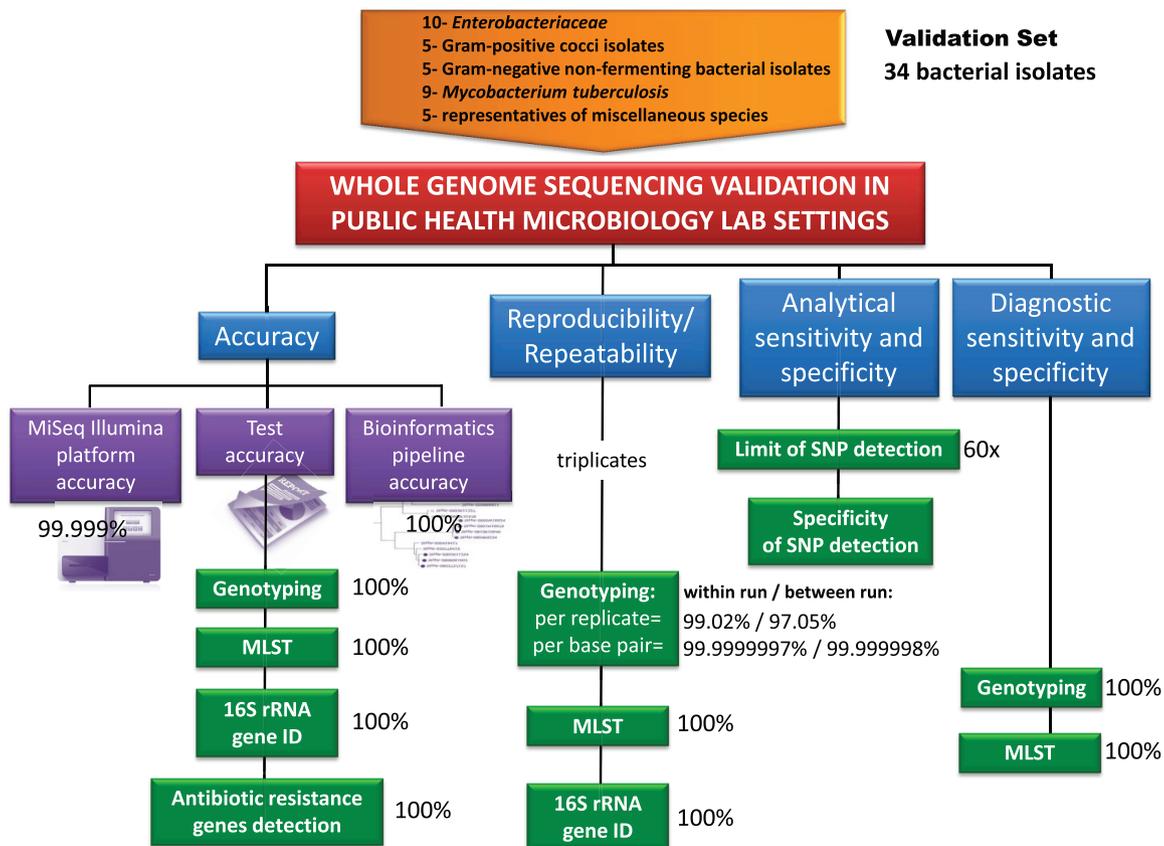
Performance characteristic for WGS applications	Definition of performance characteristic for WGS applications	Formula used for calculation	Assay-specific definition	Result of assay used for validation of parameter		
				hqSNP-based genotyping	MLST	16S
Accuracy	Degree of agreement between the nucleic acid sequences derived from the assay (measured value) and those from a reference sequence (true value)					
Accuracy of platform	Accuracy of base calling against the reference sequence; the accuracy of the platform was established as the agreement between base calling made by the MiSeq sequencer (measured value) and the NCBI/CDC reference sequence (true value)	$\% \text{ agreement with reference} = \frac{[\text{covered genome length}] - (\text{total no. of SNPs differing from reference})}{[\text{covered genome length}]} \times 100$	Accuracy of the platform	99.999378%		
Accuracy of assay	Accuracy of assay is determined as an agreement of the assay result for validation sequences generated by the PHL with the assay result for reference sequences of the same strains	$\text{Accuracy} = \frac{(\text{no. of correct results})}{(\text{total no. of results})} \times 100$	Definition of correct result		Detection and correct identification of each of the MLST alleles	ID of the 16S rRNA sequence of the validation sample matches the ID of the 16S rRNA sequence of the reference sequence
Accuracy of bioinformatics pipeline	Agreement of the clustering suggested by previous investigators with the clustering achieved by analysis using PHL validation bioinformatics pipeline	$\% \text{ agreement} = \frac{(\text{no. of outbreak isolates clustered correctly in validation tree})}{(\text{total no. of outbreak isolates clustered together in the study tree})} \times 100$	Single test unit Accuracy of assay Accuracy of bioinformatics pipeline	100% 100%	Allele 100% NA	16S rRNA ID result 100% NA
Precision	Degree to which repeated sequence analyses give the same result repeatably (within-run precision) and reproducibly (between-run precision)					Antibiotic resistance gene 100% NA
Repeatability	Repeatability was established by sequencing the same samples multiple times under the same conditions and evaluating the concordance of the assay results and performance	$\text{Repeatability} = \frac{(\text{no. of within-run replicates in agreement})}{(\text{total no. of tests performed for within-run replicates})} \times 100$	Definition of correct result	Repeatability of single nucleotide variant detection	Repeatability of allele detection	Repeatability of 16S ID
Reproducibility	Reproducibility was assessed as the consistency of the assay results and performance characteristics for the same sample sequenced under different conditions, such as between different runs, operators, and sample preparations	$\text{Reproducibility} = \frac{(\text{no. of between-run replicates in agreement})}{(\text{total no. of tests performed for between-run replicates})} \times 100$	Single test unit Repeatability Definition of correct result	SNP (precision per genome size) 99.02%, 99.999997% Repeatability of single nucleotide variant detection	Allele 100% Repeatability of allele detection	16S rRNA ID 100% Repeatability of 16S ID

(Continued on next page)

TABLE 1 (Continued)

Performance characteristic for WGS applications	Definition of performance characteristic for WGS applications	Formula used for calculation	Assay-specific definition	Result of assay used for validation of parameter			Antibiotic resistance gene detection
				hqSNP-based genotyping	MLST	16S	
Analytical sensitivity (LOD)	Minimum coverage that allows accurate SNP detection (LOD <sub>SNP</sub> )	NA	Single test unit	SNP (precision per replicate), SNP (precision per base pair)	Allele	16S rRNA ID	NA
Analytical specificity (interference)	Ability of an assay to detect only the intended target in the presence of potentially cross-reacting nucleotide sequences	NA	Reproducibility	97.05%, 99.999998%	100%	100%	NA
Diagnostic sensitivity	Likelihood that a WGS assay will detect sequence variation when present within the analyzed genomic region (this value reflects the false-negative rate of the assay)	Diagnostic sensitivity = TP / (TP + FN) × 100	LOD <sub>SNP</sub>	60×	Cross-reactivity and interference from contaminating sequencing reads are possible	NA	NA
Diagnostic specificity	Probability that a WGS assay will not detect sequence variations when none are present within the analyzed genomic region (this value reflects an assay's false-positive rate)	Diagnostic specificity = TN / (TN + FP) × 100	Definition of true-positive result	No. of validation samples that clustered together with samples genetically distant according to the reference tree	No. of correctly identified alleles	NA	NA
Reportable range	Region of the genome in which a sequence of an acceptable quality can be derived by the laboratory assay	NA	Definition of false-negative result	No. of validation samples that failed to cluster together with samples genetically similar according to the reference tree	No. of unidentified or misidentified alleles in validation samples	NA	NA
			Single test unit	Individual sample clustering	Allele	NA	NA
			Diagnostic sensitivity	100%	100%	NA	NA
			Definition of true-negative result	No clustering between unrelated samples (no. of validation samples with clustering results matching the reference)	No. of unidentified alleles in negative-control samples	NA	NA
			Definition of false-positive result	No. of validation samples that failed to cluster together with samples genetically similar according to the reference tree	No. of identified alleles in negative-control samples	NA	NA
			Single test unit	Individual sample clustering	Allele	NA	NA
			Diagnostic specificity	100%	100%	NA	NA
			Genome-wide hqSNPs	Housekeeping genes in MLST scheme	16S rRNA gene	Genes in ResFinder database	

<sup>a</sup>See details in Document S1 in the supplemental material. Abbreviations: TP, true-positive results; TN, true-negative results; FP, false-positive results; FN, false-negative results; LOD, limit of detection; LOD<sub>SNP</sub>, limit of SNP detection; ID, identification; NA, the parameter was not defined for the given assay.



**FIG 1** Summary of WGS validation. The estimated performance parameters are shown in blue boxes. The components of WGS accuracy determined in this study are shown in purple boxes. The WGS assays evaluated in order to deduce the corresponding performance parameters are shown in green boxes. Percentages alongside the boxes represent values measured during this validation for the corresponding parameters.

The accuracy of WGS was divided into three components: platform accuracy, assay accuracy, and bioinformatics pipeline accuracy.

**Platform accuracy.** Platform accuracy was assessed as the accuracy of the identification of individual base pairs (“base calling”) in the bacterial genome. The accuracy of the platform was established by determining the proximity of agreement between base callings made by the MiSeq sequencer (measured value) and the NCBI/CDC reference sequence (the true value). We determined the accuracy of the MiSeq Illumina platform by mapping generated reads to the corresponding reference sequence and identifying single nucleotide polymorphisms (SNPs) throughout the genome. A few validation samples differed from the reference genome by several SNPs. However, 99% (324 out of 327) of these SNPs were reproducible among all five replicates sequenced for each sample. Since amplification and sequencing errors were random between different library preparations, it was unlikely that the same erroneous SNP would occur in all five replicates. Therefore, we concluded that these discrepancies were not caused by sequencing errors but most likely were the result of the accumulation of mutations in the reference strains or previous sequencing mistakes in the reference sequence. Sanger sequencing confirmed the SNPs found in WGS results for the selected isolates (see Appendix 14 in Document S1 in the supplemental material). However, in several instances, it was impossible to design specific primers within reach of Sanger sequencing read lengths because SNPs were found in the repeat regions. In one case, the results of Sanger sequencing matched the results of mapping, but upon comparison with *de novo* assemblies, the same sequence was found in two variants, with and without the SNP detected by mapping. This highlighted the limitation of the short-read sequencing technology and the underlying caution that must be exercised in the interpretation of

the SNPs called in the repeat regions of the genomes. In both cases, whether we took into account all SNPs detected between validation and reference sequences or only those SNPs that did not appear in all of the replicates (true sequencing errors), we observed >99.999% agreement of the generated whole-genome sequences with the reference sequences for each tested sample.

**Assay accuracy.** Assay accuracy was defined as the agreement of the assay results for the validation sequences with the assay results for the reference sequences of the same strains. Four applications of WGS were used to validate the accuracy of the assay: an *in silico* multilocus sequence typing (MLST) assay, a 16S rRNA gene species identification assay, an assay for the detection of ABR genes, and a genotyping assay using high-quality SNPs (hqSNPs).

The definition for a correct result for MLST corresponds to the correct identification of each of the MLST alleles in the validation sequence as well as in the reference sequence analyzed in parallel. For all validation samples, each of the sequences of the seven housekeeping genes used in the typing scheme (or six genes for *Aeromonas hydrophila*) were identified correctly, resulting in 100% allele identification accuracy.

For ABR gene detection, the sequences of the ATCC susceptibility control strains generated during validation were analyzed by using ResFinder. Data analysis using ResFinder implies comparison of the sequence against each entry in the ResFinder database, which at the moment of validation contained sequences of 1,719 antibiotic resistance genes, resulting in a total of 1,719 tests performed for each validation sample. The results of ResFinder detection were compared to the resistance genes known to be present in the ATCC strains. Two of the sequenced isolates contained one resistance gene each, which were also detected by ResFinder, with no additional genes being identified. No resistance genes were detected by ResFinder in the samples suggested to be negative susceptibility controls and lacking resistance determinants. Thus, the accuracy of the assay for ABR gene detection was 100% (see the section on detection of resistance genes in ATCC strains using ResFinder in Document S1 in the supplemental material). Moreover, 13 reference sequences of Gram-negative (7 sequences) and Gram-positive (6 sequences) bacterial isolates with various resistance genes from the U.S. Food and Drug Administration (FDA)-CDC Antimicrobial Resistance Isolate (AR) Bank were used for *in silico* testing of ABR gene detection accuracy using ResFinder. The resistance genes in the AR Isolate Bank isolates were previously detected by the CDC using PCR-based methods (for primary resistance types) and by ResFinder (database last updated 2 June 2016). Our analysis of reference sequences with ResFinder (last updated 17 February 2017) confirmed the presence of all genes that were detected by PCR-based methods ( $n = 8$ ), resulting in 100% accuracy. At the moment of analysis, ResFinder did not have the ability to detect the truncation of porin genes; therefore, porin-related resistance mechanisms mentioned in the CDC database could not be detected with ResFinder. Upon ResFinder analysis at the CDC, the isolates harbored a total of 83 resistance genes (representative of 57 different alleles). The detection of the genes by the CDC using ResFinder was replicated in this study, with few discrepancies (see the section on *in silico* detection of resistance genes in FDA-CDC isolates using ResFinder and Appendix 15 in Document S1 in the supplemental material). In the case of Gram-negative bacteria, all discrepancies were represented by additional genes detected in comparison with CDC ResFinder results, and these discrepancies were most likely caused by the database update; e.g., the *aph(3')-Ia* gene was not on the list of the genes detected with ResFinder by the CDC but was detected in the same sequences using ResFinder in this study. This particular gene discrepancy could be explained by the use of a later version of ResFinder, which was updated with additional aminoglycoside resistance genes. In the case of Gram-positive bacteria, several discrepancies were also caused by the additional genes detected in this study, but two genes that were present in the CDC results were missing from our results. These false-negative genes were detected by us but possessed <99% identity or an

**TABLE 2** Summary of data from previous studies used for validation of the bioinformatics pipeline

Study parameter	Value for study	
	MRSA study <sup>a</sup>	Salmonella study <sup>b</sup>
Microorganism	Methicillin-resistant <i>Staphylococcus aureus</i>	<i>Salmonella enterica</i> serovar Typhimurium
Source of isolates	Human	Human
No. of isolates analyzed	7 outbreak isolates (1 outbreak cluster) + 2 epidemiologically unrelated isolates	9 outbreak isolates (4 outbreak clusters) + 2 epidemiologically unrelated isolates
Type of outbreak	Hospital-associated outbreak	Foodborne outbreaks
Samples used for validation	P1, P2, P3, P4, P16, P21, and P25; an isolate identified by infectious control investigation as belonging to nonoutbreak ST1; a MRSA isolate identified by searching a microbiology database as belonging to nonoutbreak ST772	0803T57157, 0808S61603, 0808F31478, 0903R11327, 0811R10987, 0804R9234, 0810R10649, 0901M16079, 0110T17035, 1005R12913, and 1006R12965
GenBank accession no. of corresponding samples	ERR070045, ERR070042, ERR070043, ERR070044, ERR124429, ERR124433, ERR128708, ERR070041, ERR072248	ERR277220, ERR277226, ERR277223, ERR277222, ERR277224, ERR277221, ERR277227, ERR277228, ERR277203, ERR277233, ERR277234
No. of clusters in the study tree	1	4
No. of clusters in the validation tree	1	4
No. of outbreak isolates in each cluster in the study tree	7 for cluster 1	2 for cluster 1, 3 for cluster 2, 2 for cluster 3, and 2 for cluster 4
No. of outbreak isolates in each cluster in the validation tree	7 for cluster 1	2 for cluster 1, 3 for cluster 2, 2 for cluster 3, and 2 for cluster 4
No. of epidemiologically unrelated isolates in the set	2	2
No. of epidemiologically unrelated isolates that clustered with outbreak isolates	0	0
% agreement $\{[(\text{no. of outbreak isolates clustered correctly in the validation tree}) \times 100]/(\text{total no. of outbreak isolates that clustered together in the study tree})\}$	$(7 \times 100/7) = 100$	$(9 \times 100/9) = 100$

<sup>a</sup>See reference 38. ST1, sequence type 1; ST772, sequence type 772.

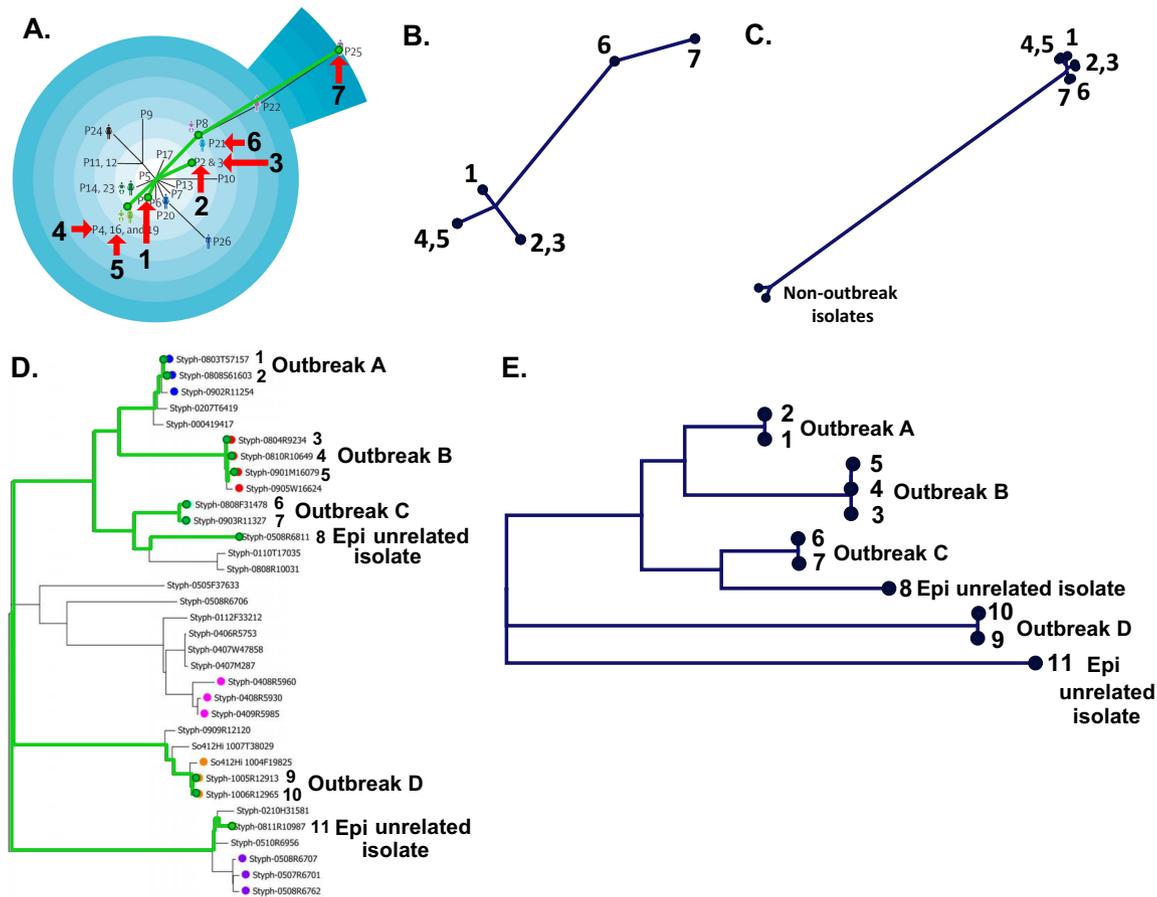
<sup>b</sup>See reference 39.

incomplete sequence and therefore were excluded from the final result. Nevertheless, the agreement between CDC ResFinder results and our ResFinder results was 99.97%.

For the 16S rRNA identification assay, variations in one gene were detected, so the species identification results as a whole (e.g., "*Escherichia coli*") were considered a single test. The identity of the 16S rRNA sequences extracted from validation samples showed 100% matches with the 16S rRNA sequences extracted from the reference sequence.

To assess the accuracy of the genome-wide hqSNP-based genotyping assay, phylogenetic trees were built by using reference sequences and validation sequences, and the resulting trees were compared. For better comparison, we used at least five strains of the same species in the phylogenetic tree. The accuracy of the genotyping assay was determined by using two approaches: (i) topological similarity between the reference tree and the validation tree using Compare2Trees software and (ii) comparison of the clustering patterns of the validation tree and the reference tree. Therefore, in addition to the detection of SNP differences between the validation sequences and reference sequences, their effect on the final tree topology was also estimated to determine the accuracy of the genotyping assay. Phylogenetic trees were generated for five bacterial species (*Escherichia coli*, *Salmonella enterica*, *Staphylococcus aureus*, *Enterococcus faecalis*, and *Stenotrophomonas maltophilia*). All five validation trees had matching clustering patterns and 100% topological similarity with the corresponding reference trees (see Table S2 in the supplemental material).

**Accuracy of the bioinformatics pipeline.** The accuracy of the bioinformatics pipeline for hqSNP-based genotyping by itself was assessed by the recapitulation of previously reported results using WGS raw reads of bacterial isolates included in two previous studies (Table 2). This was an additional assessment of the accuracy of the



**FIG 2** Bioinformatics pipeline validation with outbreak isolates from two previously published studies. (A) Phylogenetic tree of outbreak isolates reported in the “MRSA study” by Harris et al. (38). The isolates from the MRSA study that were picked for validation are indicated by arrows and numbers assigned for purposes of validation (1 to 7). (B) Phylogenetic tree validation using the samples from the MRSA study and the validation bioinformatics pipeline. The same isolates in the original tree and the validation tree are marked with the same numbers. (C) Comparison of the group of related isolates (isolates 1 to 7) from the MRSA study with epidemiologically unrelated isolates from the same study using the validation bioinformatics pipeline. (D) Phylogenetic tree combining epidemiologically related and nonrelated isolates reported in the “*Salmonella* study” by Leekitcharoenphon et al. (39). The isolates from the *Salmonella* study that were picked for validation are marked with green node circles and have the numbers 1 to 11 assigned for purposes of validation. Epi, epidemiologically. (E) Validation phylogenetic tree generated for the samples from the *Salmonella* study using the in-house bioinformatics pipeline. The same isolates in the tree from the *Salmonella* study and the validation tree are marked with the same numbers.

genotyping pipeline apart from the accuracy of base calling from in-house sequencing of the isolates. Phylogenetic analyses of isolates associated with outbreaks caused by a Gram-positive pathogen in the first study (MRSA [methicillin-resistant *Staphylococcus aureus*] study) (38) and a Gram-negative pathogen in the second study (*Salmonella* study) (39) were performed for the validation of the bioinformatics pipeline (Fig. 2). The clustering of the validation tree completely replicated the clustering of the tree from the MRSA study (38) (Fig. 2A and B); e.g., isolates 4 and 5 were identical and clustered together according to the MRSA study, and the same results were shown in the validation tree, with isolates 4 and 5 sharing the same node. All conclusions with regard to the genetic relatedness of the isolates that can be drawn from the tree reported in the MRSA study could also be made for the analysis of the corresponding validation tree. The group of outbreak isolates from the MRSA study was compared with epidemiologically unrelated isolates suggested by the same study (no tree was available from that report). Phylogenetic analysis using the in-house bioinformatics pipeline showed that epidemiologically unrelated isolates did not cluster with the group of outbreak isolates and appeared to be genetically distant (Fig. 2C). Thus, the resulting phylogenetic tree produced by our bioinformatics pipeline showed complete concordance with the epidemiological data.

From the *Salmonella* study (39), we selected nine isolates that were representative of four independent outbreaks and two isolates that were epidemiologically unrelated controls (Fig. 2D). The clustering of the validation tree was identical to the clustering of the tree from the *Salmonella* study. For example, isolates 6 and 7 were part of the same outbreak, while isolate 8 was an epidemiologically unrelated control used in that study. In agreement with the epidemiological data and the tree from the *Salmonella* study, the validation tree showed that isolates 6 and 7 clustered together but not with isolate 8 (Fig. 2E). All observations about the genetic relatedness of the isolates drawn from the tree from the *Salmonella* study could be replicated from the analysis of the validation tree. In summary, based on *in silico* analysis of data from both studies, 100% accuracy of our bioinformatics pipeline was established for phylogenetic analysis.

**WGS repeatability and reproducibility.** Repeatability (within-run precision) was established as the concordance of the assay results and quality metrics obtained for a sample tested multiple times within the same sequencing run. Reproducibility (between-run precision) was assessed as the consistency of the assay results and quality metrics for the same sample sequenced on different occasions. Thirty-four validation samples each were sequenced three times in the same sequencing run (for repeatability) and three times in different runs (for reproducibility). For within-run replicates, one DNA extract was used, but independent library preparations were done, with the final samples being included in a single sequencing run. Therefore, for each sample, the numbers of within-run replicates and between-run replicates were 3 each, and the total numbers of repeated results were 5. All quality parameters (depth of coverage, uniformity of coverage, and accuracy of base calling [Q score], etc.) did not change significantly for within- and between-run replicates, as determined by a two-tailed *t* test. For the quality values for all sequenced samples, see the section on inter- and intra-assay agreement and Appendix 8 in Document S1 in the supplemental material.

The reproducibility and repeatability of the WGS assay were evaluated with two methods: evaluation of base calling reproducibility and repeatability per replicate and evaluation of base calling reproducibility and repeatability relative to genome size. All validation samples except C50 yielded identical whole-genome sequences for all three within-run replicates. One out of three within-run replicates for isolate C50 (*Pseudomonas aeruginosa* ATCC 27853) had one SNP difference from other within-run replicates (Table S3). The repeatability per replicate was 99.02%. Three validation samples had one out of three between-run replicates that differed from the other two between-run replicates. Sample C47 (*Staphylococcus epidermidis* ATCC 12228) had one between-run replicate with two SNPs that differed from the other replicates. Samples C49 (*Streptococcus pneumoniae* ATCC 6305) and C55 (*Escherichia coli* ATCC 25922) each had one between-run replicate that differed from other replicated sequences by one SNP. The reproducibility of base calling per replicate was 97.05%. Both the reproducibility and repeatability of base calling relative to the genome size (in relation to the total number of base pairs of the covered genome size) were >99.9999%.

We also estimated the reproducibility and repeatability of MLST and 16S rRNA identification assays. For MLST, a total of 441 alleles were analyzed for within- or between-run replicates. Each single allele in all validation samples was identified consistently among within- and between-run replicates. Within- and between-run replicates had repeatable/reproducible sequences of 16S rRNA genes and resulted in the consistent identification of the species. The reproducibility and repeatability of allele detection and species identification within and between runs were 100%.

**WGS sensitivity and specificity.** The sensitivity of WGS was assessed as (i) analytical sensitivity (minimum coverage that allows accurate SNP detection) and (ii) diagnostic sensitivity (the likelihood that a WGS assay will detect sequence variation when it is present) (this value reflects the false-negative rate of the assay).

The specificity of WGS was determined as (i) analytical specificity (the ability of an assay to detect only the intended target in the presence of potentially cross-reacting nucleotide sequences) and (ii) diagnostic specificity (the probability that a WGS assay

will not detect sequence variations when none are present) (this value reflects the false-positive rate of the assay).

**Analytical sensitivity.** The limit of detection (LOD) is traditionally defined as “the lowest actual concentration of an analyte in a specimen that can be consistently detected . . . with acceptable precision” (40). The LOD in this sense is not applicable to WGS, which utilizes pure bacterial cultures as starting material; the amount of the analyte (genomic DNA) that is added to the reaction mixture is strictly standardized, and the DNA concentration of each sample is measured by a fluorometric method before each assay. In our workflow, we established the minimum amount of the starting DNA input to be 1 ng (at a concentration of 0.2 ng/ $\mu$ l) and did not process DNA extracts with concentrations of <1 ng/ $\mu$ l. Instead, we determined the LOD of SNP detection (LOD<sub>SNP</sub>) by establishing the lowest coverage that allows accurate SNP calling. The LOD<sub>SNP</sub> was estimated by modeling different mapping coverages and estimating the number of SNPs called at each of the coverage values. Nine samples representative of different species were *in silico* downsampled to coverages of 60 $\times$ , 50 $\times$ , 40 $\times$ , 30 $\times$ , 20 $\times$ , 15 $\times$ , 10 $\times$ , and 5 $\times$ . The LOD<sub>SNP</sub> was established to be 60 $\times$ , as it was the lowest coverage which yielded accurate SNP detection for all of the samples (see Table S4 and the section on analytical sensitivity of SNP detection in Document S1 in the supplemental material).

**Analytical specificity.** Analytical specificity is referred to as the “ability of an assay to detect only the intended target and that quantification of the target is not affected by cross-reactivity from related or potentially interfering nucleic acids or specimen-related conditions” (40). Since our WGS pipeline is not intended for clinical specimens directly, the conditions interfering with fragmentation, amplification, and sequencing processes are minimized and are monitored via multiple QC steps, including DNA and library purity and concentration measurements. However, interference from contaminating nucleotide sequences is much more consequential. Analytical specificity (interference) was determined by creating sequencing files containing a mixture of the reads from two different samples *in silico*, thus mimicking contamination and demonstrating its effect on mapping metrics (percentage of reads mapped, percentage of the reference sequence covered, etc.) and SNP detection (see Table S5 and the section on the analytical specificity of SNP detection in Document S1 in the supplemental material). As expected, contaminating reads led to a decrease in the percentage of mapped reads and an increase in the portion of unmapped reads. The percentage of reads in pairs decreased for samples containing contaminating reads. Modeled contamination with *E. coli* C1 and *M. tuberculosis* C57 sequencing reads did not cause any change in called SNPs. Contamination with any of the other reads led to additional SNPs being called both between the compared samples and with the reference sequence. In the sample contaminated with *S. enterica* C75, in addition to nonspecific SNPs, one of the SNPs detected previously was missing. The bioinformatics pipeline had a certain tolerance of contaminating reads depending on the nature of the contamination.

Diagnostic sensitivity and specificity were estimated for genotyping and MLST assays.

**Diagnostic sensitivity and specificity of genotyping.** To estimate the diagnostic sensitivity and specificity of WGS-based genotyping, the hqSNP phylogenetic trees generated from the validation sequences were compared to the trees generated from the reference sequences for the same strain. In the case of whole-genome sequence genotyping, the true sequence variation (SNP), which was not detected, represents a false-negative result. Changes introduced into the DNA sequence during library preparation/sequencing or data analysis errors could result in false-positive sequence variations. All generated validation trees repeated clustering and had 100% topological similarity to the corresponding reference trees, indicating the absence of false-negative or false-positive results of the genotyping assay. The hqSNP-based genotyping assay was 100% sensitive and specific.

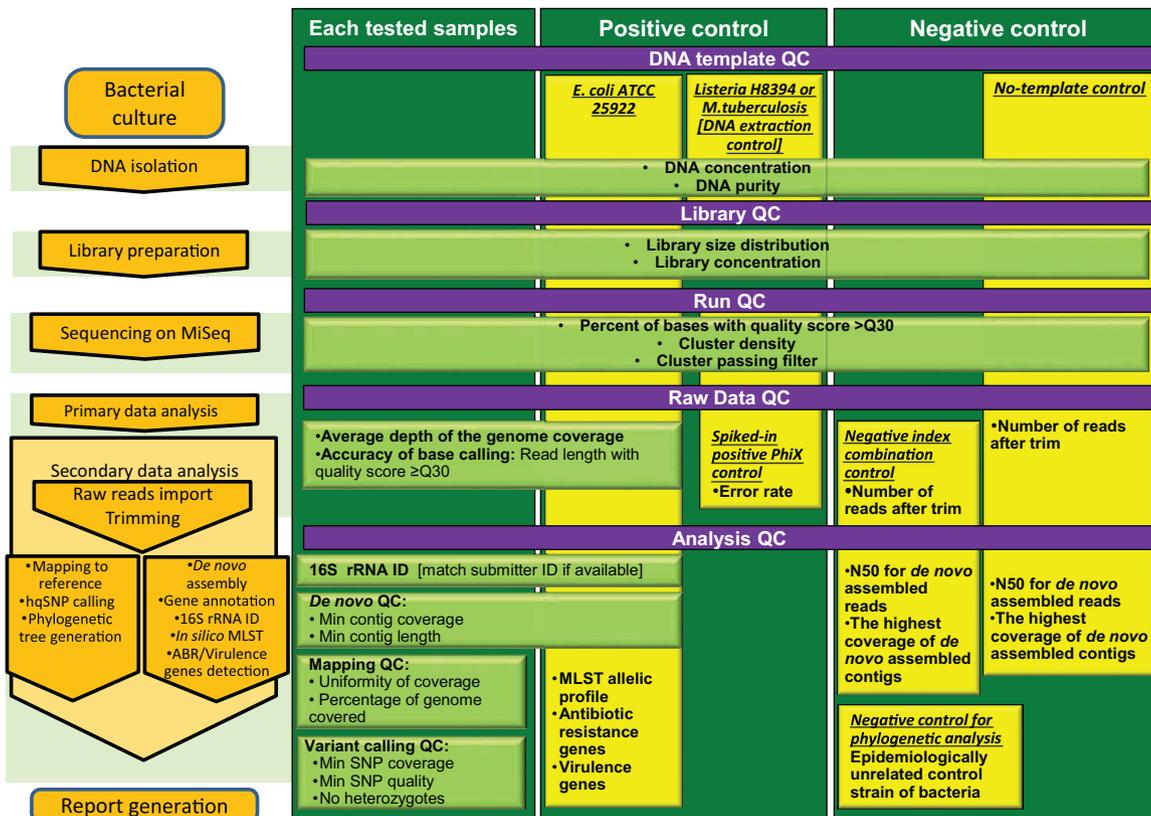
**Diagnostic sensitivity and specificity of MLST.** The sequence types of validation sequences and reference sequences were determined by using organism-specific MLST databases. For MLST, the number of true-positive results corresponds to the number of alleles correctly identified in the validation samples. For true-negative results, we performed a comparison of validation sequences against MLST databases for unmatched species, e.g., a search of alleles for the *Escherichia coli* C1 validation sample against the MLST database for *Salmonella enterica*. In the latter case, the MLST assay was expected not to identify any alleles. All alleles in the positive validation samples were identified correctly. None of the alleles were identified in the negative controls. Both the diagnostic sensitivity and analytical specificity of the *in silico* MLST assay were 100%.

**Reportable range for WGS.** The following information about the sequenced genome was collected for the reportable range: genome-wide hqSNPs, housekeeping genes used in MLST schemes, 16S rRNA genes, and antibiotic resistance genes included in the ResFinder database. A reporting language was developed to assist in the interpretation of results by an end user with or without specific WGS knowledge. The appropriate template and examples are provided in Document S1 in the supplemental material.

**Quality assurance and quality control for WGS.** QA and QC measures were developed to ensure high quality and consistency of routine testing using the MiSeq Illumina platform. QC was performed during the preanalytical (DNA isolation and library preparation), analytical (quality metrics of the sequencing run), and postanalytical (data analysis) steps of WGS. Five QC checkpoints were implemented throughout the WGS procedure: DNA template QC, library QC, sequencing run QC, raw data QC, and data analysis QC. The quality parameters evaluated at each of these QC checkpoints are summarized in Fig. 3. The QA and QC manual established for WGS applications is presented in Document S1 in the supplemental material. Based on the preliminary quality thresholds, all runs passed from the first attempt, and none of the samples had to be resequenced. The final quality thresholds were set at the lower border of quality values, which still allowed the generation of accurate and reproducible assay results (phylogenetic analysis, MLST, 16S identification, or ABR detection). The WGS quality cutoff values are summarized in Table S6 in the supplemental material. We determined the optimal depth of coverage to be  $\geq 60\times$  based on the accuracy of SNP detection at various simulated genome coverages. Table S6 does not reflect the optimal data quality parameters but merely the thresholds below which data will be rejected and the sample will have to be resequenced.

We employed two levels of positive and negative controls. First, as an internal positive control, we used spiked-in PhiX Control (Illumina Inc., San Diego, CA, USA) and evaluated its sequencing error rate. As an internal negative control, at the demultiplexing step, we called index combinations, which did not correspond to any samples in the current sequencing run but were used in the previous run. This negative-index-combination control allowed us to capture carryover contamination with the library fragments generated in the previous runs. Second, as an external positive control, we included *E. coli* strain ATCC 25922 processed from the DNA extraction step all the way to the data analysis step. As an external negative control, a no-template control was processed through all WGS steps starting with DNA extraction. Additional positive controls representing one Gram-positive strain and one *M. tuberculosis* strain were introduced solely for the DNA extraction step in cases when the corresponding types of samples were processed (to control for differences in the DNA extraction protocols). The quality cutoff values for the controls can be found in Table S6.

**Validation summary.** Analytical sensitivity (LOD of SNP detection) was established at a  $60\times$  depth of genome coverage. Analytical specificity in the presence of interfering sequencing reads was demonstrated for different types of contamination. The WGS assay was shown to have  $>99.9\%$  accuracy,  $>99.9\%$  reproducibility/repeatability, and



**FIG 3** WGS quality control scheme. The preanalytical, analytical, and postanalytical steps of WGS are shown on the left in orange boxes. The QC metrics that are being evaluated at each step are presented in light green boxes. Three vertical blocks represent the types of samples to which major QC metrics are being applied: test samples, positive controls, and negative controls. The overlap between the boxes designating QC metrics and the vertical blocks shows which QC metrics are applicable to which of the three types of samples.

100% diagnostic specificity and sensitivity. These parameters met the CLIA requirements for laboratory-developed tests (LDTs).

**DISCUSSION**

This study established the workflow and reference materials for the validation of WGS for routine use in PHLs according to CLIA guidelines for LDTs. The validation panel, sequencing analytics, and raw sequences generated during this study could serve as resources for future multilaboratory comparisons of WGS data. Additionally, the WGS performance specifications and modular validation template developed in this study could be easily adopted for the validation of other platforms and reagent kits. These results could strengthen the concept of unified laboratory standards for WGS enunciated by some professional organizations, including the Global Microbial Identifier (GMI) initiative (30, 31, 33, 41). A few other groups have also highlighted the challenges and solutions for the implementation of WGS in clinical and public health microbiology laboratories (21, 42).

Using a combination of reference strains and corresponding publicly available genomes, we devised a framework of “best practices” for the quality management of the integrated “wet-lab” and “dry-lab” WGS workflows (“pipeline”). The importance of reference materials for the validation and QC of wet- and dry-lab WGS processes was noted previously (28, 31, 33). Unlike human genomics (43), there is no well-established resource of reference materials for the validation of WGS in public health microbiology laboratories. The main challenge for creating a customized validation set is a lack of reference materials, in other words, strains that can be easily acquired by PHLs and that have high-quality, well-characterized reference genomes available. While the use of genomic sequences of ATCC strains from the NCBI is an option, it is far from being

perfect. The genome sequences available from public databases are generated by using different methods, chemistries, and platforms, which might yield different error rates; therefore, deposited sequences are not guaranteed to be free of such errors. With the ongoing development of new sequencing technologies and improvements in the quality of sequences, it is likely that the genomes sequenced with old methods might appear less accurate than the sequences generated by the laboratory during validation. Additionally, there is the possibility of mutation accumulation in control strains, e.g., ATCC cultures, after many rounds of subculture in different laboratories. Overall, there is no gold standard available yet for use as reference material for the validation of WGS for pathogenic bacteria. Nevertheless, NCBI, ENA, and similar public genome depositories remain the best resources for the genomic sequences of control strains that could be used for validation, such as the FDA Database for Regulatory-Grade Microbial Sequences (FDA-ARGOS) (<https://www.ncbi.nlm.nih.gov/bioproject/231221>). In the future, it would be helpful to have a network/agency/bank that could distribute panels of sequenced and curated isolates with genomic sequences available online for WGS validation. In the absence of such a resource, we developed a validation set of microorganisms, which can be used for future validations of WGS platforms. Bacterial genomes vary in size, GC content, the abundance of repeat regions, and other properties, which affect WGS results. We created a validation set that reflects the diversity of microorganisms with various genome sizes and GC contents that are routinely sequenced by PHLs. Different species of Gram-positive and Gram-negative microorganisms and *M. tuberculosis* were included to account for the differences in DNA extraction procedures as well.

Samples were validated based on four core elements reflected in the formal assay report: 16S rRNA-based species identification, *in silico* MLST, hqSNP phylogenetic analysis, and the presence of ABR determinants. Specific WGS assays should be validated, in addition to platform accuracy, to account for the ability to reach decent coverage in genomic areas of interest and tolerate certain base-calling error rates. Overall, we achieved high accuracy, reproducibility, repeatability, diagnostic sensitivity, and specificity for all assay analytes ranging from 99 to 100%, which exceeds the 90% threshold for LDT performance parameters per CLIA requirements. These findings are in agreement with several recent reports of 93% to 100% accuracies in WGS identification, subtyping, and antimicrobial resistance gene detection for a number of pathogens (44–47). We determined SNP detection at coverages of 5× to 60× and established an LOD for SNP detection at 60×, which was the lowest coverage that yielded accurate SNP detection in all of the samples. We determined the effect of contaminating reads on the analytical specificity of WGS. We point out that CLIA LDT performance parameters are difficult to apply to WGS analysis. For example, in determining the accuracy of the platform, CLIA would allow for up to 10% of base calls to be incorrect, which, in the case of the ~5-Mb *E. coli* genome, would mean 500,000 inaccurate SNPs, which is clearly an unacceptable error rate for any WGS application. Genome-wide hqSNP detection can be used as a way to validate platform accuracy since it allows assessment of the accuracy of base calling throughout the genome. Additionally, validation of the actual hqSNP genotyping pipeline would be required, as the phylogenetic assay takes into account not only the number of SNPs detected between isolates but also how it affects the tree topology, because one erroneous SNP is unlikely to change genotyping conclusions in most instances.

Successful CLIA integration for WGS would also require a laboratory to implement a continuous performance measurement plan via an internal or external PT program. Such PT programs are under active development, with the GMI network, the Genetic Testing Reference Materials Coordination Program (Get-RM), the Genome in a Bottle (GIAB) Consortium, and CDC PulseNet NextGen being the most prominent (31, 43). A set of generic standards has been proposed by the CAP molecular pathology checklist (30). The proposed quality standards include both live cultures as well as “sequence-only” formats for a comprehensive assessment of the WGS pipeline. Our validation set of isolates is amenable to both internal and external quality assurance testing. In

preliminary internal PT, we were able to successfully assess the entire workflow and personnel performance (details not shown).

Microbial WGS remains a dynamic technology, and therefore, any validated pipeline is unlikely to remain static. For this reason, the implementation of a modular validation template becomes crucial for the seamless and timely introduction of changes to the pipeline; e.g., we had to carry out several amendments to the protocol since its implementation in the laboratory. These amendments included a new processing algorithm for highly contagious pathogens and minor adjustments in the data analysis algorithm. The changes were accomplished via minor modifications of the pipeline with corroborative testing using a modular validation template developed by us. We also performed a two-sequencer agreement study to allow the processing of increased volumes of samples (see Document S1 in the supplemental material). The WGS report format could pose challenges for the end user. The report format in our study was designed to convey assay results to an end user with or without extensive knowledge of WGS. Additional disclaimers were used to avoid erroneous interpretations of results, for example, the disclaimer that the detection of antibiotic resistance genes by WGS does not guarantee the resistance of the strain *in vivo* and that phenotypic susceptibility testing is required to confirm antimicrobial resistance (see section on results reporting in Document S1 in the supplemental material).

This study possesses certain limitations. First, only a limited number of WGS-based assays were included in the validation study based on the most common PHL applications. Other types of WGS assays/analytics would have to be validated in a similar manner to determine the performance specifications, which are required to generate accurate and reproducible results, e.g., a threshold for the base-calling accuracy of the platform or a depth of coverage of specific genes. Second, not all validation set samples had available NCBI database entries to provide comparison sets. Third, the absence of any eukaryotic pathogens in the current validation scheme would require the implementation of a specialized pipeline for pathogenic fungi and parasites. Finally, the presence of extrachromosomal elements or polymorphic or repeat regions can negatively affect reference mapping and the *de novo* assembly of the reads generated on short-read platforms such as the Illumina platform; however, the direct effect of such genomic features on assay performance was not investigated in this study.

As the clinical and public microbiology community implements high-quality WGS, it would be opportune to consider the available models for the delivery of these services (48). Since their inception, most WGS activities have taken place in reference facilities with rather large supporting infrastructures. Although inevitable in the early stages, the centralization of services presents several challenges, such as turnaround time and access to specific expertise on the local population structure of a given pathogen, which is crucial for the management of infectious diseases at the local and regional levels. WGS services could be delivered locally and more easily with affordable sequencers, standardized reagents, and well-defined quality metrics. The local-delivery model would also be more responsive to the needs of the target clients and enhance the adoption of WGS across health care systems. Another alternative is a hybrid model with complementary central and local services to balance the need for speed with advanced expertise and resources (48). Two prominent examples of hybrid models in the United States are the FDA GenomeTrakr network for the tracking of foodborne pathogens and the CDC Advanced Molecular Detection (AMD) initiative for the improved surveillance of infectious diseases (49, 50). The AMD and GenomeTrakr frameworks rely on a participatory model with enhanced analysis, curation, and data storage at a central site. However, these resource-intensive networks focus on few selected pathogens at present. Notably, there are still significant challenges for the implementation of comprehensive WGS services at the local level (42, 51). We hope that the quality framework proposed in the present study will advance the localization of comprehensive WGS services in clinical and public health laboratories.

In summary, the salient achievements of this study included (i) the establishment of performance specifications for WGS applications in PHLs according to CLIA guidelines,

**TABLE 3** List of strains used for validation and corresponding reference materials<sup>a</sup>

MDL ID	Species	Reference genome	
		NCBI strain	NCBI accession no.
C1	<i>Escherichia coli</i> O157:H7 CDC EDL 933	<b>O157:H7 CDC EDL 933</b>	<b>NZ_CP008957.1</b>
C3	<i>Escherichia coli</i> ATCC 8739	<b>ATCC 8739</b>	<b>NC_010468.1</b>
C55	<i>Escherichia coli</i> ATCC 25922	<b>ATCC 25922</b>	<b>NZ_CP009072.1</b>
C4	<i>Enterobacter cloacae</i> ATCC 13047	<b>ATCC 13047</b>	<b>NC_014121</b>
C6	<i>Salmonella enterica</i> serovar Typhimurium ATCC 14028	<b>14028S</b>	<b>NC_016856</b>
C5	<i>Staphylococcus aureus</i> ATCC 25923	<b>ATCC 25923</b>	<b>NZ_CP009361</b>
C46	<i>Enterococcus faecalis</i> ATCC 29212	<b>ATCC 29212</b>	<b>NZ_CP008816</b>
C47	<i>Staphylococcus epidermidis</i> ATCC 12228	<b>ATCC 12228</b>	<b>NC_004461</b>
C48	<i>Staphylococcus saprophyticus</i> ATCC 15305	<b>ATCC 15305</b>	<b>NC_007350</b>
C49	<i>Streptococcus pneumoniae</i> ATCC 6305	ATCC 700669	FM211187
C50	<i>Pseudomonas aeruginosa</i> ATCC 27853	FRD1	NZ_CP010555
C51	<i>Stenotrophomonas maltophilia</i> ATCC 13637	<b>ATCC 13637</b>	<b>NZ_CP008838</b>
C52	<i>Legionella pneumophila</i> SG-12 ATCC 43290	<b>ATCC 43290</b>	<b>NC_016811</b>
C53	<i>Moraxella catarrhalis</i> 87A-3084	ATCC 25240	NZ_CP008804
C54	<i>Acinetobacter baumannii</i> ATCC 17945	PKAB07	NZ_CP006963
C103	<i>Bacteroides fragilis</i> ATCC 25285	638R	NC_016776
C104	<i>Haemophilus influenzae</i> ATCC 10211	KR494	NC_022356
C2	<i>Aeromonas hydrophila</i> ATCC 7966	<b>ATCC 7966</b>	<b>NC_008570</b>
C105	<i>Corynebacterium jeikeium</i> ATCC 43734	<b>ATCC 43734</b>	<b>GG700813, GG700833</b>
C106	<i>Neisseria gonorrhoeae</i> ATCC 49226	MS11	NC_022240
C56	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C57	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C58	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C59	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C61	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C65	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C67	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C68	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3
C69	<i>Mycobacterium tuberculosis</i>	H37Rv	NC_000962.3

<sup>a</sup>Boldface type indicates reference strains for which genomes are available from the NCBI database. Lightface type indicates cases where the genome is not available from the NCBI database and an alternative reference genome was used for mapping. MDL, Microbial Diseases Laboratory, California Department of Public Health, Richmond, CA.

(ii) development of quality assurance and quality control measures, (iii) a reporting format for end users with or without WGS expertise, (iv) availability of a validation set of microorganisms to be used for future validations, and (v) creation of a modular template for the validation of WGS processes in PHLs.

## MATERIALS AND METHODS

**Bacterial isolates and sequences.** A set of 34 bacterial isolates representing the typical workflow in PHLs was used for validation and quality control of WGS. These isolates included 10 *Enterobacteriaceae* isolates, 5 Gram-positive bacterial pathogens, 5 Gram-negative nonfermenting bacterial pathogens, 9 *Mycobacterium tuberculosis* isolates, and 5 miscellaneous bacterial pathogens (Tables 3 and 4). The bacterial pathogens for the validation set were selected to represent various genome sizes and GC contents and to account for differences in the DNA extraction protocols. We selected ATCC strains with

**TABLE 4** List of strains used for validation and corresponding reference materials available from the CDC<sup>a</sup>

MDL ID	Species	Reference raw reads generated by the CDC		Reference genome used for mapping	
		CDC strain	GenBank accession no.	NCBI strain	NCBI accession no.
C72	<i>Escherichia coli</i> O121:H19	<b>2014C-3857</b>	<b>SRR1610033</b>	2011C-3493	NC_018658
C73	<i>Salmonella enterica</i> serovar Enteritidis	<b>CDC_2010K-1543</b>	<b>SRR518749</b>	P125109	NC_011294.1
C74	<i>Salmonella enterica</i> serovar Infantis	<b>2014K-0434</b>	<b>SRR1616809</b>	1326/28	NZ_LN649235
C75	<i>Salmonella enterica</i> serovar Adelaide	<b>2014K-0941</b>	<b>SRR1686419</b>	P125109	NC_011294.1
C76	<i>Salmonella enterica</i> serovar Worthington	<b>2012K-1219</b>	<b>SRR1614868</b>	P125109	NC_011294.1
C77 <sup>b</sup>	<i>Salmonella enterica</i> serovar Saintpaul	<b>2014K-0875</b>	<b>SRR1640105</b>	14028S	NC_016856

<sup>a</sup>Boldface type indicates the reference strains for which genomes are available from the NCBI database. MDL, Microbial Diseases Laboratory, California Department of Public Health, Richmond, CA.

<sup>b</sup>Sample C77 was sequenced by the MDL only for genotyping assay accuracy validation. No replicates were done.

whole-genome sequences available from the NCBI; the NCBI sequences served as references for the validation sequencing results. Several isolates with reference sequences available from the CDC were also included in the validation set.

**Reference whole genomes.** The genome sequences of ATCC strains, isolates characterized by the CDC, and other representative isolates were downloaded from the NCBI database (<http://www.ncbi.nlm.nih.gov/genome/>) to be used as a reference according to CLSI guidelines (37) (Tables 3 and 4).

**WGS wet-bench workflow.** Whole-genome sequencing was performed on an Illumina MiSeq sequencer (see Fig. S1 in the supplemental material). The Nextera XT library preparation procedure and 2 × 300-cycle MiSeq sequencing kits were used (Illumina Inc., San Diego, CA, USA). Illumina Nextera XT indexes were used for barcoding. Bacterial DNA was extracted by using the Wizard Genomic DNA kit (Promega, Madison, WI, USA). The bacterial DNA concentrations were measured by using Qubit fluorometric quantitation with a Qubit double-stranded RNA (dsDNA) BR assay kit (Thermo Fisher Scientific, Waltham, MA, USA). DNA purity was estimated by using a NanoDrop 2000 UV-visible (UV-Vis) spectrophotometer (NanoDrop Products, Wilmington, DE, USA). The Mastercycler nexus was used for incubation for fragmentation reaction and PCR (Eppendorf North America, Hauppauge, NY, USA). The library concentration was measured by using a Qubit HS kit. The DNA library size distribution was estimated by using a 2100 BioAnalyzer instrument and a High Sensitivity DNA analysis kit (Agilent Technologies, Santa Clara, CA, USA). Ampure beads were used for size selection. Manual normalization of libraries was performed. Genome sizes of samples pooled in one run were taken into account to ensure the equal representation of each genome per flow cell. The number of samples was kept at 16 per run (except for one run with 19 samples), resulting in a total genome load of 48 to 84 Mb/run (average, 69 Mb/run). The PhiX Control V3 sequencing control at 1% was spiked in for every sequencing run (Illumina Inc., San Diego, CA, USA). Genomes were generated with a depth of coverage (estimated for trimmed mapped reads) in the range of 15.71× to 216.4× (average, 79.72×; median, 71.55×).

The preliminary run and raw data acceptance criteria were set for (i) the percentage of bases with a ≥Q30 quality score for the run of ≥50%, with the requirement that the Q30 score for the generated genome sequences must be ≥75% for at least 80 bp of the read length; (ii) a minimum average coverage of the genome of 10×; (iii) a PhiX error rate that must be <6%; and (iv) the negative control meeting the following parameters: <10,000 reads after trimming, an N50 value of <1,000, and the highest coverage of *de novo*-assembled contigs of <10×. Data which did not meet these quality parameters were rejected. The preliminary quality thresholds were further adjusted based on the quality ranges observed during validation.

**Bioinformatics pipeline.** Paired-end reads were quality trimmed with the threshold of Q30 and then used for mapping to the reference and *de novo* assemblies on CLCbio Genomic Workbench 8.0.2 (Qiagen, Aarhus, Denmark). The BAM files generated after mapping to the reference genome were taken through series of software suites to generate the phylogenetic tree. A customized shell script was created to automate the subsequent steps after mapping, which included (i) calculating the genotype likelihood using SAMtools mpileup (v.1.2) (52), (ii) conversion to a VCF matrix using bcftools view (v0.1.19; <http://samtools.github.io/bcftools/>), (iii) single nucleotide polymorphism calling in coding and noncoding genome areas using bcftools call —c (v0.1.19), (iv) variant parsing using vcftools (v0.1.12b) (53) to include only hqSNPs with a mean coverage of ≥30× and a minimum SNP quality (minQ) of ≥200 with indels and heterozygote calls excluded and the allele frequency defining a homozygous call at ≥93%, and (v) conversion of the SNP matrix to a FASTA alignment file for export back to CLCbio Genomic Workbench 8.0.2 for the generation of the phylogenetic tree.

**hqSNP-based genotyping.** Maximum likelihood phylogenetic trees were generated based on hqSNPs under the Jukes-Cantor nucleotide substitution model; bootstrapping included 100 replicates.

**16S rRNA gene-based identification.** Genomes were annotated with prokka v1.1 (54), and species identification was performed by comparing 16S rRNA gene sequences against data in the Ribosomal Database Project (RDP) database (55).

**In silico MLST.** *In silico* MLST was performed by using the Center for Genomic Epidemiology (CGE) online tool (56).

**Detection of antibiotic resistance genes.** ABR gene detection was performed by using the CGE ResFinder online resource (57) with 99% identity and 100% query length coverage thresholds. Two sets of sequences were analyzed. For the first set, ATCC reference bacterial strains designated for use as antibiotic susceptibility controls were sequenced by the laboratory performing the validation. Negative controls were chosen among strains described as being susceptible, with no known antibiotic resistance genes, according to CLSI document M100-S25 (58). Positive controls were chosen among strains that possessed resistance determinants, according to CLSI document M100-S25. For the second set, reference sequences were acquired from the FDA-CDC Antimicrobial Resistance Isolate Bank (<https://www.cdc.gov/drugresistance/resistance-bank/>) for *in silico* testing. Thirteen isolates (7 Gram-negative and 6 Gram-positive organisms) with various resistance genes were analyzed.

**Validation plan.** Thirty-four bacterial isolates were sequenced in triplicate. The WGS pipeline, including wet- and dry-bench processes, was validated by assessing the performances of the platform, specific WGS-based assays, and the bioinformatics analysis pipeline (Fig. 1). Validated WGS assays included genome-wide SNP-based genotyping, MLST, 16S rRNA species identification, and antibiotic resistance gene detection. The following performance characteristics were assessed: accuracy, reproducibility (between-run precision), repeatability (within-run precision), analytical and diagnostic sensitivity, and analytical and diagnostic specificity.

For reproducibility assessments, all between-run replicates were generated starting from fresh cultures except for *M. tuberculosis*, where DNA samples were used. Between-run replicates were pro-

cessed on three different days, alternating between two operators, as recommended in CLSI document MM11A (59). For repeatability, within-run replicates were prepared from one DNA extract, but independent library preparations were done, with the final samples being included in one sequencing run. The reproducibility and repeatability of the genotyping assay were evaluated with two methods: (i) evaluation of the reproducibility and repeatability of SNP calling per replicate takes into account the whole genome of one replicate as a single test, meaning that any number of single nucleotide changes in one out of three of the replicates was regarded as 33.3% disagreement for that validation sample, and (ii) the reproducibility and repeatability of SNP calling relative to the genome size were calculated by taking a discordant SNP into account as a percentage of the number of base pairs in the genome (genome size); e.g., a 10-SNP difference among the replicates with 5-Mb genomes equals 99.9998% reproducibility.

**Analytical sensitivity.** The LOD of SNP calling was estimated by modeling different mapping coverages and estimating the minimum coverage that allowed accurate SNP calling ( $LOD_{SNP}$ ). Mapping BAM files were downsampled in order to achieve different coverage values (60×, 50×, 40×, 30×, 20×, 15×, 10×, and 5×) for each of the nine samples representative of the different species. The original sequence mapping coverage was estimated from the BAM file by using the following command: `samtools depth bamfile_sorted.bam | awk '{sum+= $3} END {print "Average =",sum/NR}'`. The bam files were downsampled to the desired coverage by using the command `samtools view -h -s F bamfile_sorted.bam > bamfile_sorted_30x.bam`, where F is a fraction of the desired coverage in relation to the original coverage. Downsampled BAM files were compared to other replicates for the same sample at their original coverage.

**Analytical specificity.** To determine analytical specificity (interference), we generated *in silico* sequencing files containing a mixture of reads from two different samples, thus mimicking contamination. The effect of potentially interfering sequencing reads on mapping metrics (percentage of reads mapped/not mapped, percentage of the reference sequence covered, etc.) and SNP detection was estimated. Sample C3 (*Escherichia coli* ATCC 8739) was selected as a "control, not-contaminated sample," and equal parts of reads from different species were merged with it to generate mixed fastq files. This was done by using the concatenate command, e.g., `cat file1 file2 > mixed file`. To ensure that the correct file was generated, the numbers of lines in both files and the mixed file were counted. The number of lines in the mixed file was the sum of the numbers of lines in the two parent files. Also, the column headings were checked to ensure that the headers of the mixed file were similar to those of the parent file. "Contaminated" and "not-contaminated" reads were mapped to the same reference genome of *E. coli* ATCC 8739 from the NCBI. Further SNP calling analysis was performed by using the standardized pipeline under validation. The specificity of SNP calling in the samples containing potentially interfering reads was estimated by comparing (i) the number of SNPs detected between the contaminated sequence and the reference sequence and (ii) the number of SNPs detected between the contaminated and not-contaminated sequences included in the same SNP calling analysis (vcf file after filtering).

**Accession number(s).** Data from this whole-genome shotgun project have been deposited in GenBank under raw read accession numbers [SRR4114366](https://www.ncbi.nlm.nih.gov/seq/submit/SRR4114366) to [SRR4114399](https://www.ncbi.nlm.nih.gov/seq/submit/SRR4114399) and assembly accession numbers [MTFS00000000](https://www.ncbi.nlm.nih.gov/seq/submit/MTFS00000000) to [MTGZ00000000](https://www.ncbi.nlm.nih.gov/seq/submit/MTGZ00000000) (see Table S1 in the supplemental material). All WGS data, including sequences of the replicates, are associated with BioProject accession number [PRJNA341407](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA341407).

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/JCM.00361-17>.

**SUPPLEMENTAL FILE 1**, PDF file, 12.7 MB.

## ACKNOWLEDGMENTS

This study was supported in part by grant 6NU50CK000410-03 from the HHS, U.S. Centers for Disease Control and Prevention.

We are thankful to Eija Trees from the Centers for Disease Control and Prevention for assistance in acquiring reference sequences for the comparative study.

V.K.K. and V.C. were responsible for conceptualization, data analysis, and the original draft; V.K.K., C.-L.T., A.L.G., and J.C. performed data collection and analysis; R.M. performed software coding; V.K.K., C.-L.T., R.M., A.L.G., J.C., and V.C. reviewed the draft; and V.C. was responsible for project administration and supervision.

## REFERENCES

- Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. 2012. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* 13:601–612. <https://doi.org/10.1038/nrg3226>.
- Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 8:e1002824. <https://doi.org/10.1371/journal.ppat.1002824>.
- Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, DeSmet T, Gujja S, McCowan C, Montmayeur A, Steelman S,

- Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill F-X, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci U S A* 109:3065–3070. <https://doi.org/10.1073/pnas.1121491109>.
4. McGann P, Bunin JL, Snesrud E, Singh S, Maybank R, Ong AC, Kwak YI, Seronello S, Clifford RJ, Hinkle M, Yamada S, Barnhill J, Lesho E. 2016. Real time application of whole genome sequencing for outbreak investigation—what is an achievable turnaround time? *Diagn Microbiol Infect Dis* 85:277–282. <https://doi.org/10.1016/j.diagmicrobio.2016.04.020>.
  5. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TEA, Crook DW, Walker AS, Golubchik T. 2014. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol* 52:1182–1191. <https://doi.org/10.1128/JCM.03117-13>.
  6. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *J Clin Microbiol* 52:2365–2370. <https://doi.org/10.1128/JCM.00262-14>.
  7. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17:333–351. <https://doi.org/10.1038/nrg.2016.49>.
  8. Roetzler A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsche-Gerdes S. 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10:e1001387. <https://doi.org/10.1371/journal.pmed.1001387>.
  9. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364:730–739. <https://doi.org/10.1056/NEJMoa1003176>.
  10. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TEA, Walker AS. 2013. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med* 369:1195–1205. <https://doi.org/10.1056/NEJMoa1216064>.
  11. Etienne KA, Roe CC, Smith RM, Vallabhaneni S, Duarte C, Escandón P, Castañeda E, Gómez BL, de Bedout C, López LF, Salas V, Hederra LM, Fernández J, Pidal P, Hormazabal JC, Otaíza-O'Ryan F, Vannberg FO, Gillette J, Lemmer D, Driebe EM, Engelthaler DM, Litvintseva AP. 2016. Whole-genome sequencing to determine origin of multinational outbreak of *Sarocladium kilienae* bloodstream infections. *Emerg Infect Dis* 22:476–481. <https://doi.org/10.3201/eid2203.151193>.
  12. Pallen MJ. 2014. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology* 141:1856–1862. <https://doi.org/10.1017/S0031182014000134>.
  13. Onderdonk AB, Delaney ML, Fichorova RN. 2016. The human microbiome during bacterial vaginosis. *Clin Microbiol Rev* 29:223–238. <https://doi.org/10.1128/CMR.00075-15>.
  14. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, Kliethermes S, Schreckenberger PC, Brubaker L, Gai X, Wolfe AJ. 2014. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* 5:e01283-14. <https://doi.org/10.1128/mBio.01283-14>.
  15. Schubert AM, Rogers MAM, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, Schloss PD. 2014. Microbiome data distinguish patients with *Clostridium difficile* infection and non-*C. difficile*-associated diarrhea from healthy controls. *mBio* 5:e01021-14. <https://doi.org/10.1128/mBio.01021-14>.
  16. Naccache SN, Federman S, Veerarahgavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk K-C, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Girard G, Leroy E, Schneider BS, Fair JN, Martinez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J, Miller S, Chiu CY. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* 24:1180–1192. <https://doi.org/10.1101/gr.171934.113>.
  17. Hasman H, Saputra D, Sichert-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. 2014. Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol* 52:139–146. <https://doi.org/10.1128/JCM.02452-13>.
  18. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZM, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M. 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxicogenic *Escherichia coli* O104:H4. *JAMA* 309:1502–1510. <https://doi.org/10.1001/jama.2013.3231>.
  19. Endrullat C, Glöckler J, Franke P, Frohme M. 2016. Standardization and quality management in next-generation sequencing. *Appl Transl Genom* 10:2–9. <https://doi.org/10.1016/j.atg.2016.06.001>.
  20. Salto-Tellez M, Gonzalez de Castro D. 2014. Next-generation sequencing: a change of paradigm in molecular diagnostic validation. *J Pathol* 234:5–10. <https://doi.org/10.1002/path.4365>.
  21. Goldberg B, Sichtig H, Geyer C, Ledebner N, Weinstock GM. 2015. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *mBio* 6:e01888-15. <https://doi.org/10.1128/mBio.01888-15>.
  22. Kwong JC, McCallum N, Sintchenko V, Howden BP. 2015. Whole genome sequencing in clinical and public health microbiology. *Pathology* 47:199–210. <https://doi.org/10.1097/PAT.0000000000000235>.
  23. Luheshi LM, Raza S, Peacock SJ. 2015. Moving pathogen genomics out of the lab and into the clinic: what will it take? *Genome Med* 7:132. <https://doi.org/10.1186/s13073-015-0254-z>.
  24. Oliver GR, Hart SN, Klee EW. 2015. Bioinformatics for clinical next generation sequencing. *Clin Chem* 61:124–135. <https://doi.org/10.1373/clinchem.2014.224360>.
  25. Fricke WF, Rasko DA. 2014. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* 15:49–55. <https://doi.org/10.1038/nrg3624>.
  26. Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, Rusu LI. 2014. WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare? *Pathogens* 3:437–458. <https://doi.org/10.3390/pathogens3020437>.
  27. Rhoads DD, Sintchenko V, Rauch CA, Pantanowitz L. 2014. Clinical microbiology informatics. *Clin Microbiol Rev* 27:1025–1047. <https://doi.org/10.1128/CMR.00049-14>.
  28. Gargis AS, Kalman L, Lubin IM. 2016. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J Clin Microbiol* 54:2857–2865. <https://doi.org/10.1128/JCM.00949-16>.
  29. Lefterova MI, Suarez CJ, Banaei N, Pinsky BA. 2015. Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. *J Mol Diagn* 17:623–634. <https://doi.org/10.1016/j.jmoldx.2015.07.004>.
  30. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DGB, Merker JD, Nagarajan R, Palicki LA, Robetorye RS, Schrijver I, Weck KE, Voelkerding KV. 2014. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* 139:481–493. <https://doi.org/10.5858/arpa.2014-0250-CP>.
  31. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS. 2015. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect Dis* 15:174. <https://doi.org/10.1186/s12879-015-0902-3>.
  32. Olson ND, Jackson SA, Lin NJ. 2015. Report from the Standards for Pathogen Identification via Next-Generation Sequencing (SPIN) Workshop. *Stand Genomic Sci* 10:119. <https://doi.org/10.1186/s40793-015-0112-z>.
  33. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauser BA. 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* 30:1033–1036. <https://doi.org/10.1038/nbt.2403>.
  34. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, Gowrisankar S, Hegde MR, Kulkarni S, Mason CE, Nagarajan R, Voelkerding KV, Worthey EA, Aziz N, Barnes J, Bennett SF, Bisht H, Church DM, Dimitrova Z, Gargis SR, Hafez N, Hambuch T, Hyland FC, Luna RA, MacCannell D, Mann T, McCluskey MR, McDaniel TK, Ganova-Raeva LM, Rehm HL, Reid J, Campo DS, Resnick RB, Ridge PG, Salit ML, Skums P, Wong LJ, Zehnbauser BA, Zook JM, Lubin IM. 2015. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* 33:689–693. <https://doi.org/10.1038/nbt.3237>.
  35. Duncavage EJ, Abel HJ, Merker JD, Bodner JB, Zhao Q, Voelkerding KV, Pfeifer JD. 2016. A model study of in silico proficiency testing for clinical

- next-generation sequencing. *Arch Pathol Lab Med* 140:1085–1091. <https://doi.org/10.5858/arpa.2016-0194-CP>.
36. Mellmann A, Andersen PS, Bletz S, Friedrich AW, Kohl TA, Lilje B, Niemann S, Prior K, Rossen JW, Harmsen D. 2017. High interlaboratory reproducibility and accuracy of next-generation-sequencing-based bacterial genotyping in a ring trial. *J Clin Microbiol* 55:908–913. <https://doi.org/10.1128/JCM.02242-16>.
  37. CLSI. 2014. Nucleic acid sequencing methods in diagnostic laboratory medicine: approved guideline, 2nd ed. MM09-A2. CLSI, Wayne, PA.
  38. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ. 2013. Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis* 13:130–136. [https://doi.org/10.1016/S1473-3099\(12\)70268-2](https://doi.org/10.1016/S1473-3099(12)70268-2).
  39. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS One* 9:e87991. <https://doi.org/10.1371/journal.pone.0087991>.
  40. Burd EM. 2010. Validation of laboratory-developed molecular assays for infectious diseases. *Clin Microbiol Rev* 23:550–576. <https://doi.org/10.1128/CMR.00074-09>.
  41. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405–423. <https://doi.org/10.1038/gim.2015.30>.
  42. Lesho E, Clifford R, Onmus-Leone F, Appalla L, Snesrud E, Kwak Y, Ong A, Maybank R, Waterman P, Rohrbeck P, Julius M, Roth A, Martinez J, Nielsen L, Steele E, McGann P, Hinkle M. 2016. The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLoS One* 11:e0155770. <https://doi.org/10.1371/journal.pone.0155770>.
  43. Kalman LV, Datta V, Williams M, Zook JM, Salit ML, Han J-Y. 2016. Development and characterization of reference materials for genetic testing: focus on public partnerships. *Ann Lab Med* 36:513–520. <https://doi.org/10.3343/alm.2016.36.6.513>.
  44. Lindsey RL, Pousee H, Chen JC, Strockbine NA, Carleton HA. 2016. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing *Escherichia coli* (STEC) in the United States. *Front Microbiol* 7:766. <https://doi.org/10.3389/fmicb.2016.00766>.
  45. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordworth S, Wyllie D, Xu L, Crook DW. 2016. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 4:49–58. [https://doi.org/10.1016/S2213-2600\(15\)00466-X](https://doi.org/10.1016/S2213-2600(15)00466-X).
  46. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden R, Drobniewski FA, Allix-Béguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA. 2015. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 15:1193–1202. [https://doi.org/10.1016/S1473-3099\(15\)00062-6](https://doi.org/10.1016/S1473-3099(15)00062-6).
  47. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 67:2640–2644. <https://doi.org/10.1093/jac/dks261>.
  48. Arnold C. 2016. Considerations in centralizing whole genome sequencing for microbiology in a public health setting. *Expert Rev Mol Diagn* 16:619–621. <https://doi.org/10.1586/14737159.2016.1164039>.
  49. Auffray C, Caulfield T, Griffin JL, Khoury MJ, Lupski JR, Schwab M. 2016. From genomic medicine to precision medicine: highlights of 2015. *Genome Med* 8:12. <https://doi.org/10.1186/s13073-016-0265-4>.
  50. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R. 2016. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J Clin Microbiol* 54:1975–1983. <https://doi.org/10.1128/JCM.00081-16>.
  51. Robilotti E, Kamboj M. 2015. Integration of whole-genome sequencing into infection control practices: the potential and the hurdles. *J Clin Microbiol* 53:1054–1055. <https://doi.org/10.1128/JCM.00349-15>.
  52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
  53. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genome Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>.
  54. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
  55. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633–D642. <https://doi.org/10.1093/nar/gkt1244>.
  56. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 50:1355–1361. <https://doi.org/10.1128/JCM.06094-11>.
  57. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 52:1501–1510. <https://doi.org/10.1128/JCM.03617-13>.
  58. CLSI. 2015. Performance standards for antimicrobial susceptibility testing; twenty-fifth informational supplement. CLSI document M100-S25. CLSI, Wayne, PA.
  59. CLSI. 2007. Molecular methods for bacterial strain typing; approved guideline, MM11-A. CLSI, Wayne, PA.