Title

Validation, Implementation, and Clinical Utility of Whole Genome Sequence-Based Bacterial Identification in the Clinical Microbiology Laboratory.

Journal

The Journal of molecular diagnostics : JMD, 23(11)

ISSN

1525-1578

Authors

Price, Travis K
Realegeno, Susan
Mirasol, Ruel
et al.

Publication Date

2021-11-01

DOI

10.1016/j.jmoldx.2021.07.020

Peer reviewed

The Journal of Molecular Diagnostics, Vol. 23, No. 11, November 2021

the Journal of Molecular Diagnostics
jmdjournal.org

See related Guest Editorial on page 1415

Validation, Implementation, and Clinical Utility of Whole Genome Sequence-Based Bacterial Identification in the Clinical Microbiology Laboratory

Travis K. Price, Susan Realegeno, Ruel Mirasol, Allison Tsan, Sukantha Chandrasekaran, Omai B. Garner, and Shangxin Yang

From the Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California

Accepted for publication July 22, 2021.

Address correspondence to Shangxin Yang, Ph.D., D.(A.B.M.M.), Department of Pathology and Laboratory Medicine, University of California Los Angeles, 11633 San Vicente Blvd., Los Angeles, CA 90095. E-mail: shangxinyang@mednet.ucla.edu.

The application of next-generation sequencing extends from microbial identification to epidemiologic insight and antimicrobial resistance prediction. Despite this potential, the roadblock for clinical laboratories lies in implementation and validation of such complex technology and data analysis. Herein, a validation study used whole-genome sequencing (WGS) for pan-bacterial identification (ID) in a clinical laboratory, and assessed its clinical relevance. A diverse set of 125 bacterial isolates, including a subset without genus (25) and/or species (10) ID, were analyzed by de novo assembly and reference genome mapping. The 16S rRNA, rpoB, and groEL genes were used for ID. Using WGS, 100% (89 of 89) and 89% (79 of 89) of isolates were identified at the genus and species-levels, respectively. WGS also provided improved results for 71% (25/35) isolates originally reported with genus-only or descriptive IDs. Chart review identified cases in which improved genus and/or species-level ID by WGS may have had a positive impact on patient care. Reasons included the use of an ineffective antibiotic owing to unclear ID, use of antibiotics when not clinically indicated, and help with an outbreak investigation. The implementation of next-generation sequencing in a clinical microbiology setting is a challenging but necessary task. This study provides a model for the validation and implementation of bacterial ID by WGS in such a setting. *(J Mol Diagn 2021, 23: 1468–1477; https://doi.org/10.1016/j.jmoldx.2021.07.020)*

It has been almost 2 decades since the development of next-generation sequencing (NGS)-based methodologies. Since then, it has become clear that this technology inevitably will alter every aspect of the fields of clinical and public health microbiology.[1–3] The clinical and diagnostic utility is vast and includes speciation/classification of microorganisms, bacterial and viral strain typing,[4–7] metagenomic and microbiome-based investigations,[8–11] antimicrobial resistance prediction,[12–15] infection prevention surveillance,[16–18] and outbreak tracking.[19–22] Mounting evidence in the literature illustrates the benefit of each of these points. Despite the overwhelming evidence of the value and transformative potential of NGS, most clinical microbiology laboratories have yet to implement this technology.

There are several hurdles for bringing NGS to the clinical laboratory.[23–27] The assay in its entirety is complex, time consuming, and requires significant training of laboratory personnel. Integration of NGS into the laboratory workflow presents additional challenges related to quality control, infrastructure, and established reference standards. For instance, implementation of standard quality metrics is not easily adaptable from typical diagnostic assays. The cost of capital equipment and bioinformatic and computing software, although decreasing, remains relatively high. Finally, standardized guidelines for NGS are in their infancy for pathology laboratories but are essentially nonexistent for clinical microbiology. Taken together, bringing NGS into the clinical laboratory requires advanced skills, time, financial cost, flexibility, and innovation.

Supported by the University of California Los Angeles Department of Pathology and Laboratory Medicine.

T.K.P. and S.R. contributed equally to this work.

Disclosures: None declared.

https://doi.org/10.1016/j.jmoldx.2021.07.020

An NGS-based assay using whole-genome sequencing (WGS) of pure microbial isolates for bacterial identification was implemented at the University of California Los Angeles (UCLA) Clinical Microbiology Laboratory. Described here are the wet laboratory procedures, bioinformatic analyses, and various assay quality metrics used. An extensive validation study was performed to measure precision and accuracy/concordance. Included in the assay are a set of microorganisms that could not be speciated by standard techniques to assess the overall clinical utility. Finally, considerations and challenges faced in bringing this technology to the laboratory are discussed.

Materials and Methods

Bacterial Isolates and Identification Methods

One hundred twenty-six bacterial isolates, representing a variety of clinically relevant taxonomic groups, obtained from clinical specimens or reference collections, were used. Clinical isolates (116 of 126) collected from 2015 to 2019 and stored at -70°C, were retrieved and analyzed retrospectively. Reference isolates (10 of 126) were obtained from the ATCC (Manassas, VA). A description of these isolates including information regarding source, original taxonomic/descriptive identification (ID), identification method(s) used, and WGS taxonomic ID are found in Supplemental Table S1.

At the original time of isolation, clinical isolates were given taxonomic or descriptive ID using one or more of the following methods: cellular staining methods (eg, Gram, Ziehl-Neelsen), biochemical tests (eg, oxidase, catalase), commercial ID methods [ie, VITEK MS (bioMérieux, Hazelwood, MO), VITEK 2 (bioMérieux), API (bioMérieux), PathoDX (Remel, Lenexa, KS), and AccuProbe (Hologic, San Diego, CA)] and Sanger sequencing (ie, 16S rRNA, *rpoB* gene). For Sanger sequencing, the MicroSeq 500 16S rDNA-based ID system (Applied Biosystems, Carlsbad, CA), which targets the V1 to V2 region of the 16S rRNA gene (positions 1 to 527), or consensus PCR primers, were used to amplify a 764-bp (positions 2573 to 3337) variable region of the *rpoB* gene.[28] Sequences were queried using the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) nucleotide collection (nr/nt) database (updated March 2020), with program selection of highly similar sequences (megablast) (*https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome*, last accessed: July 22, 2021). Matches with 99% or higher coverage and identity were assigned species-level IDs.

A reference strain of *Escherichia coli* (ATCC 8739) was used as a positive control for each sequencing run. The remaining 125 isolates were divided into three categoric groups, which are referenced in the analysis (Supplemental Table S1). The bacterial validation set (50 of 125) included a variety of nonmycobacterial isolates with both genus and species-level IDs. The mycobacterial validation set (40 of 125) included a variety of *Mycobacterium* spp. identified at the species-level (24 of 40) or complex-level (16 of 40) (ie, *Mycobacterium avium* complex). The challenge set (35 of 125) included isolates with genus-only (25 of 35) or descriptive-only (10 of 35) IDs.

Bacterial Isolate Preparation and DNA Extraction

Mycobacterial isolates were heat-inactivated (100°C for 30 minutes) and an additional bead-beating step was performed for mechanical disruption of the cell wall. The Qiagen (Valencia, CA) EZ1 Blood and Tissue Kit and the EZ1 Advanced XL instrument were used according to the manufacturer's instructions to extract genomic DNA from pure microbial isolates. Extracted DNA was quantified with the Qubit 1× double-stranded DNA HS assay using the Qubit 3.0 Fluorometer (Thermo Fisher, Waltham, MA). Acceptable quantities of DNA were 0.04 ng/μL or more. *E. coli* DNA and PCR-grade water were used as positive and negative controls, respectively.

Library Preparation and Sequencing

DNA was diluted in water to obtain optimal concentrations within the range of 1 to 500 ng in 30 μL. Library preparation was performed using the Illumina DNA Prep Kit (Illumina, San Diego, CA) according to the manufacturer's instructions. In brief, tagmented DNA was amplified using a limited-cycle PCR for barcoding and library cleanup was performed using a two-step bead purification procedure. Libraries were quantified, as described in the previous paragraph, using Qubit. After quantification, the Agilent DNA 1000 Kit and the Agilent 2100 Bioanalyzer instrument (Agilent, Carpinteria, CA) were used to analyze the average band size of the libraries with an acceptable range of 300 to 900 bp. Molarity was calculated for each sample using library concentration and average library bp size.

Libraries with sufficient concentrations were normalized to 2 nmol/L or 4 nmol/L concentrations for the Illumina MiSeq v2 or v3 reagent kit, respectively. Equal volumes of each library were pooled and denatured for sequencing. The Illumina MiSeq System was used to produce 250-bp paired-end reads regardless of whether the v2 or v3 kits were used.

Bioinformatic Analysis

An overview of the analytical workflow is shown in Figure 1. Data were uploaded to the Illumina BaseSpace cloud and demultiplexed. Various sequence run quality control (QC) metrics were recorded: the percentage passing filter (PF), evenness of reads PF (ie, ≥2% PF reads per sample), percentage of undetermined reads, lane density, and % over Q30. Results and passing criteria are listed in Supplemental Table S2. Some metrics (ie, percentage of

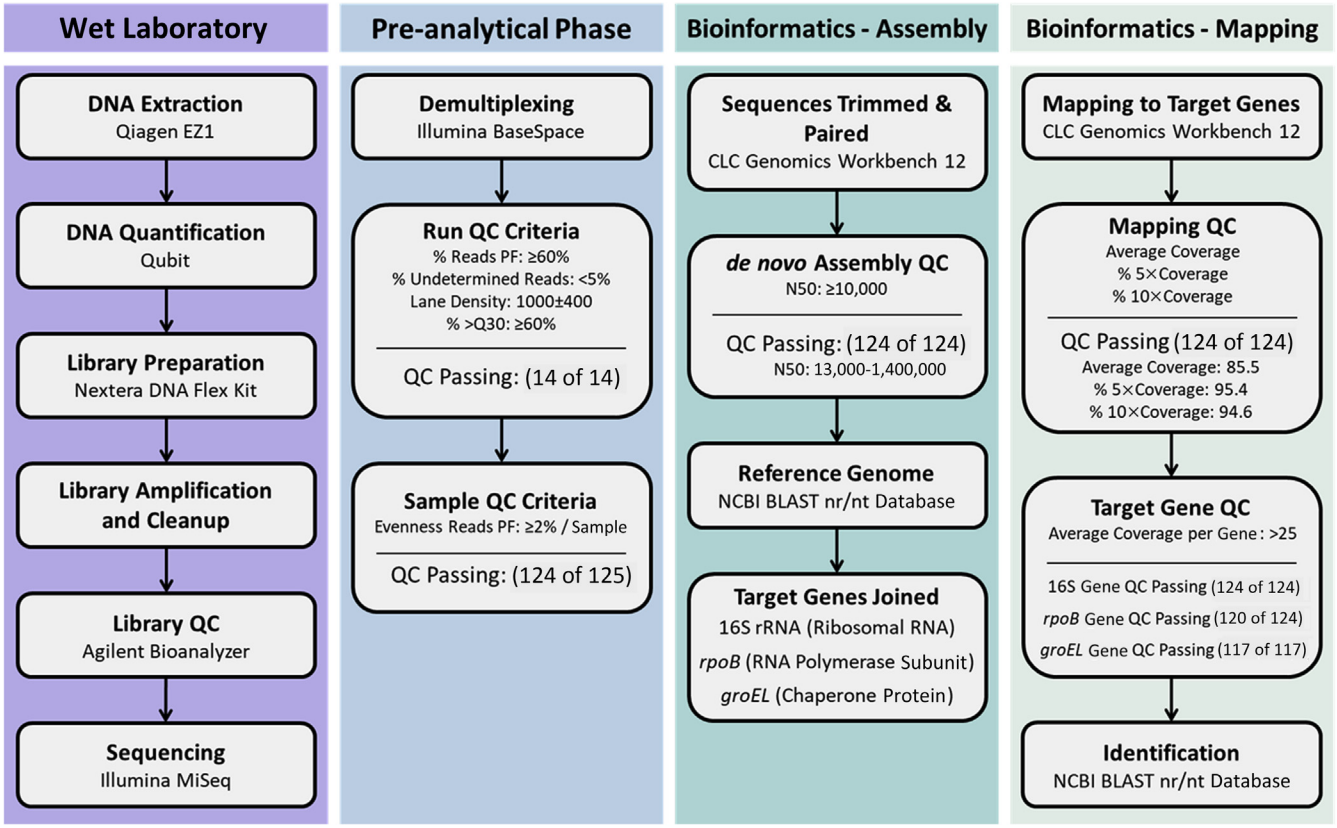

Figure 1 Chart of key quality control (QC) metrics. Nextera DNA Flex Kit (Illumina, La Jolla, CA). BLAST, Basic Local Alignment Search Tool; NCBI, National Center for Biotechnology Information; PF, passing filter.

reads PF and evenness of reads PF) were considered mandatory, meaning the sequencing run was to be rejected if the QC criterion was not met. Other metrics were considered optimal, meaning these were monitored closely and the NGS board would decide how to proceed if the QC was out of the optimal range. The percentage of reads PF for each run were recorded and tracked over time; runs with a percentage of reads PF of between 60% and 80% were reviewed by the NGS board before further data analysis.

Sequences were uploaded to the CLC Genomics Workbench version 12 software (Qiagen) where they were trimmed and paired. *De novo* sequence assembly was performed; minimum contig length was set to 1000. Various assembly QC metrics were recorded: the number of contigs, N50, and the percentage of GC content. Results are listed in Supplemental Table S3. Several contigs with a size range of 2000 to 10,000 bp were extracted and analyzed using the BLAST nr/nt database to obtain an appropriate reference genome for mapping. The chosen reference genome had 80% or more coverage and identity to the contig; these values were recorded.

The reference genome was uploaded to CLC Genomics Workbench 12 and three target genes (ie, 16S rRNA, *rpoB*, *groEL*) were identified, extracted, and concatenated. The bp size of each gene was recorded. The sequence reads then were mapped to the concatenated reference sequence and various QC metrics were recorded: the percentage of mapped reads, average coverage, and the percentage with 5× and 10× coverage. The total reads, average coverage, and percentage of ambiguous nucleotides per target gene also were recorded. Results are listed in Supplemental Table S3. The workflow used for mapping is shown in Supplemental Figure S1. Overall, minimum thresholds for QC metrics were determined based on overall analysis of the QC values collected for all samples sequenced and categorized as mandatory or optimal criteria.

The consensus sequence of each target gene was queried using the BLAST nr/nt database. The NCBI accession ID of all results with 99% or more coverage and the identity for each target gene were recorded, along with the corresponding values. If no result met this criteria, the single top result was recorded. Using the scheme shown in Figure 2, an ID was given, and the method used was recorded. If the coverage and identity for a given analysis was lower than 91%, consultation with an NGS board member (an assigned individual to review analysis) was used to determine the final ID. In one case when all three target genes had BLAST results with more than 1 identical match, a k-mer tree was generated using CLC Genomics Workbench v12 by incorporating the reference genomes of multiple highly genetically related species. The species ID ultimately was

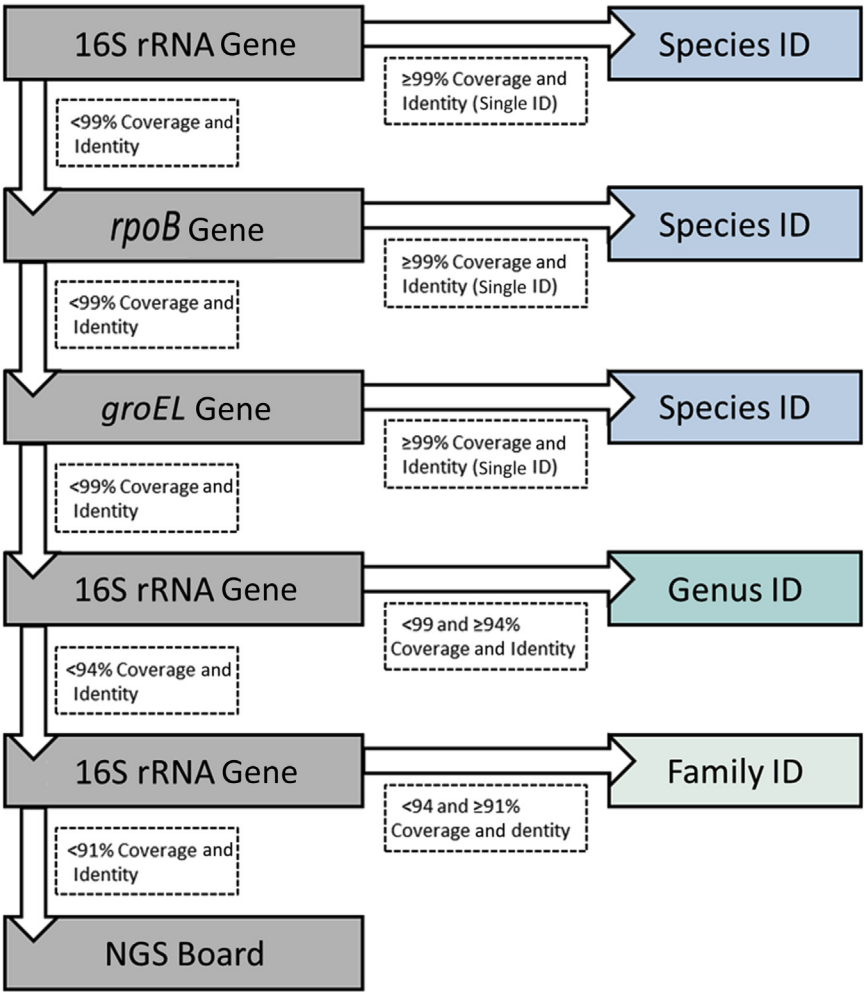

Figure 2 Cascading whole-genome sequencing (WGS)-based scheme for phylogenetic identification (ID). NGS, next-generation sequencing.

determined by the same cluster of the reference genome and the sample in the k-mer tree.

Validation

WGS-based IDs were compared with the original ID to determine concordance. The method used for the original ID was recorded and considered in subsequent discrepant analyses. Precision studies were performed with positive and negative controls. An *in silico* validation of the BLAST nr/nt database was performed. Clinical utility was assessed by chart review of patients for select clinical isolates.

Results

Quality Control, Bioinformatics Performance, and Database Validation

Fourteen sequencing runs were performed by several laboratory staff members. Various metrics were obtained from the Illumina BaseSpace Sequence Hub (Supplemental Table S2). QC criteria were established for the percentage PF and the evenness of reads PF, and were considered mandatory. These metrics and the corresponding study data are shown in Figure 1. One isolate (UCLA_082, *Mycobacterium xenopi*) was excluded because of low (<2%) evenness of reads PF. Precision studies, both within and between sequence runs, were performed using the positive and negative controls, which met our criteria (Supplemental Table S4).

Sequence-specific metrics, including *de novo* assembly and reference genome mapping, were obtained from the CLC Genomics Workbench, and the results are shown in Supplemental Table S3. QC criteria for the average gene coverage were collected and considered mandatory; data are shown in Figure 1. Four isolates did not have adequate (>25×) coverage for the *rpoB* gene, therefore all were identified using the 16S gene.

An *in silico* validation of the NCBI BLAST nr/nt database was performed using 22 complete genomes of various microorganisms downloaded from GenBank. These genomes were processed through our bioinformatics pipeline and all had concordant IDs (Supplemental Table S5).

Assay Performance

Using WGS, 99% (123 of 124) and 79% (98 of 124) of isolates were identified at the genus and species levels, respectively. Fifty-six percent (69 of 124), 59% (73 of 124), and 25% (29 of 117) of the 16S, *rpoB*, and *groEL* genes provided database matches with 99% or more coverage and identity, respectively. Raw data are shown in Supplemental Table S6. For isolates with multiple gene targets meeting our criteria (45 of 124), all IDs agreed.

Using our cascading ID scheme (Figure 2), the 16S gene was used for 70% (69 of 98) of species-level IDs; the *rpoB* gene was used for 29% (28 of 98), of which 27 were part of the mycobacterial validation set. The NGS board was used for one species-level ID. In this case, all three target genes had BLAST results with more than one identical match. The case was reviewed, and a species ID ultimately was decided upon using k-mer tree–based phylogenetic comparisons. These data are described according to the validation set in Table 1.

Concordance Results and Discordant Analysis

Bacterial and Mycobacterial Validation Sets

After one isolate (UCLA_082) was excluded because of low reads, a total of 89 isolates were included in the bacterial (50 isolates) and mycobacterial (39 isolates) validation sets (Supplemental Table S1). Using WGS, 100% and 89% (79 of 89) (bacterial set, 42 of 50; mycobacterial set, 37 of 39) of isolates were identified to the genus and species levels, respectively (Table 1).

All genus IDs were concordant (89 of 89). Twenty isolates had discordant species IDs (bacterial set, 15; mycobacterial set, 5). Of these, 10 were discordant because they did not receive a species-level ID using WGS owing to a lack of BLAST results with 99% or more coverage and identity for any of the target genes. The remaining 10 isolates were discordant because of a discrepancy between the WGS ID and the original ID. These data, along with the method used to obtain the original ID, are summarized in Table 2.

An evaluation of the discordant species IDs (20 of 89) was performed using various approaches, including reassessment of the original method used for identification (Table 2). Raw data are shown in Supplemental Table S7. For isolates in which the original ID was obtained using matrix-assisted laser desorption ionization time-of-flight (7 of 20), the isolate was re-analyzed using the most current VITEK MS database v3.2.0 (method A); this resolved two of seven discordant IDs (ie, two *Elizabethkingia meningoseptica*) (Table 2). For isolates in which the original ID was obtained using 16S Sanger sequencing (11 of 20), it was determined if the discordance was a result of a sequencing error, or a result of the difference in resolution using a full-length 16S (ie, WGS data) versus partial (ie, 16S Sanger sequencing). The V1 to V2 region (bases 1 to 527) of the

Table 1 Description of WGS ID and target gene statistics by validation set.

	Full Validation Set (N = 124)	Bacterial Validation Set (N = 50)	Challenge Set (N = 35)		Mycobacterial Validation Set (N = 39)
			Genus-only ID (N = 25)	Descriptive ID (N = 10)	
WGS ID					
Genus ID	123 (99%)	50 (100%)	24 (96%)	10 (100%)	39 (100%)
Species ID	98 (79%)	42 (84%)	16 (64%)	3 (30%)	37 (95%)
Target Gene					
16S (≥99%)	69 (56%)	42 (84%)	16 (64%)*	2 (20%)	9 (23%)
16S (≥94%)	93 (75%)	50 (100%)	24 (96%)	10 (100%)	9 (23%)
16S (≥91%)	94 (76%)	50 (100%)	25 (100%)	10 (100%)	9 (23%)
rpoB (≥99%)	73 (59%)	28 (56%)	9 (38%)	3 (30%)	33 (85%)
groEL (≥99%)	29/117 (25%)	23/48 (48%)	3/20 (15%)	3/10 (30%)	0/39 (0%)
Multiple Target Genes ≥99%	N = 45	N = 28	N = 9	N = 3	N = 5
ID Agreement	45 (100%)	28 (100%)	9 (100%)	3 (100%)	5 (100%)
Gene Used for Species ID†	N = 98	N = 42	N = 16	N = 3	N = 37
16S	69 (70%)	42 (100%)	16 (100%)	2 (67%)	9 (24%)
rpoB	28 (29%)	0 (0%)	0 (0%)	1 (33%)	27 (73%)
groEL	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NGS Board	1 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)

*Isolate UCLA_060 (*Psychrobacter* spp.) had a Basic Local Alignment Search Tool nr/nt database match of ≥99% coverage and identity for the 16S rRNA gene. However, the database result was *Psychrobacter* spp. No species ID was given, but this isolate still was included in the statistics for 16S (≥99%). This was not considered an improved result.

†Target genes with Basic Local Alignment Search Tool nr/nt database (*https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome*) matches containing multiple genus and/or species IDs with 99% or more coverage and identity were not used in the cascading ID scheme. This applied to the 16S results for 26 isolates in the mycobacterial validation set; *rpoB* ultimately was used to identify all 26 isolates. These 26 isolates were not included in the statistics for 16S (≥99%, ≥94%, ≥91%) under the Target Gene section.

ID, identification; NGS, next-generation sequencing; UCLA, University of California Los Angeles; WGS, whole-genome sequencing.

Table 2 List of Discordant Results and Adjusted Concordance Data

Set*	Study ID	Cause of discordance	WGS ID	Original ID	Method used for original ID	Resolution method
B	UCLA_088	Discrepancy	*Elizabethkingia anophelis*	*Elizabethkingia meningoseptica*	MALDI-TOF MS	A
B	UCLA_089	Discrepancy	*E. anophelis*	*E. meningoseptica*	MALDI-TOF MS	A
C	UCLA_111	Discrepancy	*Robinsoniella peoriensis*	*Clostridium* spp.	MALDI-TOF MS	A
B	UCLA_010	<99% match	*Dysgonomonas* spp. most closely related to *Dysgonomonas capnocytophagoides*	*D. capnocytophagoides*	16S Sanger sequencing	B
B	UCLA_011	<99% match	*Turicibacter* spp. most closely related to *Turicibacter sanguinis*	*T. sanguinis*	16S Sanger sequencing	B
B	UCLA_037	Discrepancy	*Tsukamurella tyrosinosolvens*	*Tsukamurella pulmonis*	16S Sanger sequencing	B
B	UCLA_050	Discrepancy	*T. tyrosinosolvens*	*T. pulmonis*	16S Sanger sequencing	B
B	UCLA_100	<99% match	*Inquilinus* spp. most closely related to *Inquilinus limosus*	*I. limosus*	16S Sanger sequencing	B
C	UCLA_099	<94% match	Peptoniphilaceae most closely related to *Anaerococcus*	*Anaerococcus* spp.	16S Sanger sequencing	B
B	UCLA_061	<99% match	*Nocardia* spp. most closely related to *Nocardia flavorosea*	*Nocardia carnea*	16S Sanger sequencing	C
B	UCLA_073	<99% match	*Legionella* spp. most closely related to *Legionella anisa*	*Legionella bozemanii*	16S Sanger sequencing	C
B	UCLA_074	<99% match	*Legionella* spp. most closely related to *Legionella donaldsonii*	*Legionella feeleii*	16S Sanger sequencing	C
B	UCLA_096	<99% match	*Actinomyces* spp. most closely related to *Actinomyces odonotolyticus*	*Actinomyces turicensis*	MALDI-TOF MS	C
B	UCLA_110	<99% match	*Actinomyces* spp. most closely related to *Actinomyces gaoshouyii*	*Actinomyces denticolens*	MALDI-TOF MS	C
M	UCLA_007	<99% match	*Mycobacterium* spp. most closely related to *M. fortuitum*	*Mycobacterium mucogenicum*	MALDI-TOF MS	D
M	UCLA_026	Discrepancy	*Mycobacterium houstonense*	*M. mucogenicum*	MALDI-TOF MS	D
M	UCLA_031	Discrepancy	*Mycobacterium alsense*	*Mycobacterium asiaticum*	MALDI-TOF MS	D
B	UCLA_020	Discrepancy	*Nocardia arthritidis*	*Nocardia beijingensis*	16S Sanger sequencing	Not resolved
B	UCLA_062	Discrepancy	*Nocardia wallacei*	*Nocardia transvalensis*	16S Sanger sequencing	Not resolved
B	UCLA_064	Discrepancy	*Leifsonia xyli*	*Leifsonia shinshuensis*	16S Sanger sequencing	Not resolved
M	UCLA_005	Discrepancy	*Mycobacterium dioxanotrophicus*	*Mycobacterium canariasense*	*rpoB* Sanger sequencing	Not resolved
M	UCLA_080	<99% match	*Mycobacterium* spp. most closely related to *Mycobacterium peregrinum*	*M. peregrinum*	ATCC	Not resolved

Method A. The original isolate was regrown and analyzed on MALDI-TOF MS using the VITEK MS database v3.2.0 in February, 2020.

Method B. The original isolate was identified using 16S Sanger sequencing of the V1 to V2 region. Basic Local Alignment Search Tool analysis of this region from the WGS data showed a match (≥99% coverage and identity) to the original ID. This shows that the discordance is not the result of a sequencing error, but rather from the increased resolution provided by using a full 16S gene versus a partial sequence.

Method C. A reference genome for the species of the original ID was downloaded from the National Center for Biotechnology Information GenBank (*https://www.ncbi.nlm.nih.gov/genbank*). A full 16S gene was identified and aligned to the WGS data using the alignment tool in Basic Local Alignment Search Tool. The alignment did not match (<99% coverage and identity). This shows that the original ID was wrong.

Method D. A reference genome for the species of the original ID was downloaded from the National Center for Biotechnology Information GenBank. A full *rpoB* gene was identified and aligned to the WGS data using the alignment tool in Basic Local Alignment Search Tool. The alignment did not match (<99% coverage and identity). This shows that the original ID was wrong.

*Group B, bacterial validation set; group M, mycobacterial validation set; group C, challenge set.

ID, identification; MALDI-TOF, matrix-assisted laser desorption ionization time-of-flight; WGS, whole-genome sequencing.

16S gene was isolated from the WGS data and assessed (method B); this resolved 5 of 11 discordant IDs (Table 2). For the remaining unresolved discordant IDs (13 of 20), the original ID was re-analyzed by comparing the 16S (method C) or *rpoB* (method D) genes from reference genomes of the original ID, with the genes obtained by WGS, using the percentage of alignment. Several isolates (8 of 13) showed a nonmatch (ie, <99% coverage and identity), indicating that the original ID was incorrect. The remaining five isolates (ie, *Nocardia beijingensis*, *Nocardia transvalensis*, *Leifsonia shinshuensis*, *Mycobacterium canariasense*, and *Mycobacterium peregrinum*) were not resolved.

Challenge Set

Thirty-five isolates were included in the challenge set: 25 with genus-only IDs and 10 with descriptive IDs (Supplemental Table 1). Using WGS, 97% (34 of 35) and 54% (19 of 35) of isolates were identified to the genus and species levels, respectively (Table 1).

Ninety-two percent (23 of 25) of genus IDs were concordant. Using the same evaluation method described above, the two discordant IDs (ie, *Clostridium* spp. and *Anaerococcus* spp.) were resolved (Table 2), resulting in an adjusted genus concordance of 100%.

Improved Results and Clinical Utility

One hundred percent (16 of 16) of *M. avium* complex isolates were speciated using WGS. In the challenge set, WGS provided improved species-level IDs for 60% (15 of 25) of isolates with genus-only IDs. A genus-level ID was given to all (10 of 10) isolates with descriptive IDs, and 30% (3 of 10) were speciated. These data are summarized in Table 3.

A chart review was performed to assess the potential clinical impact of the improved ID results. Clinical information was assessed to determine whether there would be a positive (ie, change or improvement in care, epidemiologic insight) or neutral (ie, unclear or no change in care) impact on patient care. I was determined that in nine cases the improved result would have had a positive impact on patient care (Table 3). Reasons included the use of an ineffective antibiotic because of unclear ID, use of antibiotics when not clinically indicated, and use in an outbreak investigation. Examples of each scenario are described below.

Ineffective antibiotics were given to a 37-year-old woman presenting with pelvic pain and hematuria. Cervical swab cultures grew *Mycoplasma* spp. and it was not possible to speciate the organism further using conventional methods. The patient was placed on azithromycin but saw no clinical improvement. Follow-up cultures were taken 3 months later that grew *Mycoplasma* spp. again. The isolate was included in this validation study as UCLA_106 and was determined to be *Mycoplasma hominis*, which is resistant to azithromycin. Therefore, speciation would have guided proper antimicrobial therapy use.

Unnecessary antibiotics were given to a 76-year-old man presenting for follow-up evaluation for a localized abscess. Previous cultures of the abscess drainage had grown *Corynebacterium amycolatum*. The patient was treated with amoxicillin. Follow-up cultures grew *Corynebacterium* spp. Assuming this was still *C. amycolatum*, the patient was switched to vancomycin. The isolate was included as UCLA_033 and determined to be *Corynebacterium jeikeium*, likely representing normal skin flora. Therefore, the patient did not need additional antibiotics because the causative organisms had already been cleared.

An outbreak investigation was performed after speciation of three *M. avium* complex isolates included in the validation (UCLA_128, UCLA_133, and UCLA_153). All three isolates were obtained from patients undergoing cardiac surgeries and grew from chest wall tissue or surgical swabs. All three isolates were initially identified as *M. avium* complex using DNA hybridization probes and were not speciated further. Through WGS, it was determined that the isolates were *Mycobacterium chimaera*, which prompted an investigation into the heating-cooling devices used during surgery because of the known international outbreak of these devices with *M. chimaera*. Further bioinformatic analyses showed that all three isolates were closely related, and the heater-cooler devices were removed.

Discussion

A WGS-based assay for pan-bacterial ID for use in the clinical microbiology laboratory was established and implemented in this study. Various quality metrics for the wet bench, sequencing, and bioinformatics procedures were established and the entire assay was validated using an extensive and diverse set of bacteria. Precision and accuracy/concordance were assessed. Finally, the assay was challenged using a set of bacteria that could not be speciated by standard techniques to assess the overall clinical utility. At the time of writing, this assay was implemented successfully into the UCLA clinical microbiology laboratory for routine clinical testing.

WGS-based assays for bacterial ID currently are performed at large reference and public health laboratories. However, implementation in a hospital setting presents several challenges not encountered by larger laboratories including expensive instrumentation and complex technical workflows. This study implemented a new NGS bench to perform a comprehensive workflow consisting of sample processing, wet bench procedures, and bioinformatic analyses. Preventative maintenance, documentation and storage of clinical isolates, documentation of QC metrics, database management, data storage, and troubleshooting also are tasks administered by the NGS bench as part of quality management. A group of clinical laboratory scientists were trained successfully on all aspects of the bench, competency was assessed, and they were given alternate proficiency testing using previously analyzed isolates. Financial cost was another significant challenge for implementation. Purchases included the Illumina MiSeq, network drive space to store data, and software to perform bioinformatic analysis (ie, CLC Genomics Workbench); where possible, other necessary equipment already were in use in the laboratory and were adapted for use in this assay. The cost to sequence one isolate is approximately $250. Overall, time was another challenging factor, and a workflow was optimized to reduce turnaround times and maximize sequencing capacity. Currently, we perform one sequencing run per week; turnaround times usually range from 3 to 7 days. Collectively, assay development, validation performance, data analysis, establishment of

Table 3 List of Improved Results and Clinical Utility

Set*	Study ID	Improvement category	WGS ID	Original ID	Source	Clinical utility
C	UCLA_059	Minor	*Burkholderia multivorans*	*Burkholderia* spp.	Sputum-expectorated	Neutral (appropriate antibiotics given)
C	UCLA_111	Minor	*Robinsoniella peoriensis*	*Clostridium* spp.	Blood	Neutral (complicated infection)
C	UCLA_033	Minor	*Corynebacterium jeikeium*	*Corynebacterium* spp.	Body fluid	Positive (normal flora, antibiotics not needed)
C	UCLA_038	Minor	*Corynebacterium aurimucosum*	*Corynebacterium* spp.	Blood, central line	Neutral (likely normal flora)
C	UCLA_009	Minor	*Desulfovibrio desulfuricans*	*Desulfovibrio* spp.	Blood, peripheral vein	Neutral (complicated infection)
C	UCLA_034	Minor	*Microbacterium oleivorans*	*Microbacterium* spp.	Blood, peripheral vein	Neutral (appropriate antibiotics given)
C	UCLA_106	Minor	*Mycoplasma hominis*	*Mycoplasma* spp.	Genital swab (cervix)	Positive (ineffective antibiotics given)
C	UCLA_048	Minor	*Nocardia asiatica*	*Nocardia* spp.	Bronchoalveolar lavage	Positive (speciation guides antibiotic treatment)
C	UCLA_087	Minor	*Staphylococcus pseudintermedius*	*Staphylococcus* spp.	Drainage, sinus	Neutral (complicated infection)
C	UCLA_049	Minor	*Streptococcus mitis*	*Streptococcus* spp.	Blood, peripheral vein	Neutral (appropriate antibiotics given)
C	UCLA_068	Minor	*Streptomyces coelicolor*	*Streptomyces* spp.	Body fluid (left knee)	Neutral (unlikely to cause infection)
C	UCLA_097	Minor	*Ureaplasma urealyticum*	*Ureaplasma* spp.	Urine (midstream)	Neutral (appropriate antibiotics given)
C	UCLA_113	Minor	*Ureaplasma urealyticum*	*Ureaplasma* spp.	Urine (midstream)	Neutral (appropriate antibiotics given)
C	UCLA_115	Minor	*Ureaplasma parvum*	*Ureaplasma* spp.	Urine (midstream)	Neutral (appropriate antibiotics given)
C	UCLA_118	Minor	*Ureaplasma parvum*	*Ureaplasma* spp.	Urine (midstream)	Neutral (appropriate antibiotics given)
C	UCLA_112	Minor	*Anaerococcus* spp.	Anaerobic GP cocci	Blood	Neutral (appropriate antibiotics given)
C	UCLA_116	Minor	*Anaerococcus* spp.	Anaerobic GP cocci	Blood	Neutral (appropriate antibiotics given)
C	UCLA_109	Major	*Brachyspira pilosicoli*	Curved rods	Blood	Positive (bloodstream infection)
C	UCLA_012	Major	*Aerococcus sanguinicola*	GP cocci (clusters)	Urine (midstream)	Positive (ineffective antibiotics given)
C	UCLA_017	Minor	*Pseudomonas* spp.	GN rod (nonfermenter)	Blood	Positive (bloodstream infection)
C	UCLA_042	Minor	*Dysgonomonas* spp.	GN rod (small)	Blood, peripheral vein	Neutral (appropriate antibiotics given)
C	UCLA_091	Minor	*Clostridium* spp.	Anaerobic GN rod	Blood	Neutral (appropriate antibiotics given)
C	UCLA_104	Minor	*Anaerococcus* spp.	Anaerobic GP cocci	Bone (toe)	Neutral (appropriate antibiotics given)
C	UCLA_036	Major	*Ruminococcus gnavus*	Anaerobic GP cocci	Body fluid (peritoneal)	Neutral (likely normal flora)
C	UCLA_105	Minor	*Hungatella* spp.	Anaerobic GP rod	Blood	Neutral (appropriate antibiotics given)
M	UCLA_014	Minor	*Mycobacterium avium*	MAC	Stool	Neutral (treatment of MAC is similar)
M	UCLA_128	Minor	*Mycobacterium chimaera*	MAC	Wound (chest)	Positive (related to potential outbreak)
M	UCLA_132	Minor	*M. avium*	MAC	Sputum-expectorated	Neutral (treatment of MAC is similar)
M	UCLA_133	Minor	*M. chimaera*	MAC	Tissue (chest)	Positive (related to potential outbreak)
M	UCLA_134	Minor	*M. avium*	MAC	Tissue (chest)	Neutral (treatment of MAC is similar)
M	UCLA_135	Minor	*Mycobacterium intracellulare*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_136	Minor	*M. avium*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_137	Minor	*M. avium*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_138	Minor	*Mycobacterium intracellulare*	MAC	Sputum-expectorated	Neutral (treatment of MAC is similar)
M	UCLA_139	Minor	*M. avium*	MAC	Blood, peripheral vein	Neutral (treatment of MAC is similar)
M	UCLA_140	Minor	*M. avium*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_141	Minor	*Mycobacterium intracellulare*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_142	Minor	*Mycobacterium intracellulare*	MAC	Sputum-expectorated	Neutral (treatment of MAC is similar)
M	UCLA_143	Minor	*M. avium*	MAC	Sputum-induced	Neutral (treatment of MAC is similar)
M	UCLA_151	Minor	*M. avium*	MAC	Bronchoalveolar lavage	Neutral (treatment of MAC is similar)
M	UCLA_153	Minor	*M. chimaera*	MAC	Surgical swab (chest)	Positive (related to potential outbreak)

*Group C, challenge set; group M, mycobacterial validation set.

ID, identification; GN, gram-negative; GP, gram-positive; MAC, *Mycobacterium avium* complex; UCLA, University of California Los Angeles; WGS, whole-genome sequencing.

appropriate quality metrics, writing of standard operating procedures, and training took approximately 1 year. In general, the lack of published NGS implementation strategies and guidelines contributed to the overall time burden because many innovations are required for the development of this process.

Considering the challenges described above, the implementation of such an assay to perform a diagnostic function

already achievable by current far-cheaper, simpler, and faster technologies may seem superfluous. However, the diagnostic and clinical utility of NGS in general far exceeds bacterial ID, and additional assays can be adapted easily to the already existing workflow. For example, once the genome sequence of a pure bacterial isolate has been obtained, one could determine the species or strain of the organism, perform antimicrobial resistance prediction, or investigate genomic relatedness among isolates for epidemiologic purposes including infection prevention, outbreak investigation, and infectious source control within a patient population. The wet bench procedures would not need to be altered for any of these additional tests. Therefore, the use of NGS in the clinical laboratory represents a unique diagnostic tool in that it does not provide a single function and is easily scalable to meet various clinical demands. The remaining challenges reside in the bioinformatics, which will be the main driver and roadblock for the expansion of clinical microbiology NGS tests.

In this study, three marker genes (ie, 16S, *rpoB*, *groEL*) were used to determine the ID of the bacterial isolate. The performance of each gene varied widely, with 56% (69 of 124), 59% (73 of 124), and 25% (29 of 117) of the 16S, *rpoB*, and *groEL* genes providing NCBI BLAST nr/nt matches with 99% or more coverage and identity, respectively. The 16S gene was most useful to speciate the isolates in the bacterial validation set (84%; 42 of 50), while the *rpoB* gene was most useful for the mycobacterial validation set (85%; 33 of 39). In many cases, the 16S gene could not differentiate among *Mycobacterium* spp. with multiple species matches at 100% coverage and identity. Therefore, the inclusion of the *rpoB* gene was invaluable to speciate *Mycobacterium* spp. and shows that Sanger sequencing of just one marker gene would not be sufficient for a pan-bacterial ID assay.

Improved IDs were provided for a total of 41 bacterial isolates and, therefore, shows the clinical utility of NGS using our assay. Two of these isolates were from critical specimens (ie, blood) in which the full microorganism ID is essential for appropriate treatment. However, there is limited practicality in this given the 1- to 2-week turnaround time for this assay. Other NGS-based metagenomics approaches would be clinically valuable in these situations. Other isolates with improved IDs showed that improper antimicrobial therapy was given to the patient, likely leading to recurrence of infection. Finally, some isolates with improved IDs (eg, *M. chimaera*) were valuable for infection-prevention purposes because they indicated a potential outbreak related to a heater-cooler device (unpublished data, Xu, 2021).

There are several important factors to consider when developing an NGS-based assay for a clinical laboratory. First, because of the complexity of the technology, there are plenty of opportunities for error. To manage this, we implemented a weekly NGS board, comprising senior members of the clinical microbiology laboratory, to review the data, troubleshoot, and discuss any technical or analytical issues. The NGS board also makes decisions on challenging cases in which additional bioinformatic analysis and literature review are required to report our results to the clinicians. Treating physicians are also invited to participate in the board meeting discussion to provide direct input about the clinical aspects of the cases. Second, as a measure of QC, it is important to correlate the NGS results with other available laboratory tests. Here, we correlated our IDs to phenotypic and other diagnostic tests performed in the laboratory. Next, 22 publicly available complete genomes from a set of diverse bacteria were used to verify the NCBI BLAST nr/nt database. Because of the reliance on a publicly available and frequently updated database, routine verification is necessary. As with other commonly used laboratory methods, such as matrix-assisted laser desorption ionization time-of-flight mass spectrometry, the NGS result is heavily reliant on the database used. Third, it is important to document quality metrics for each part of the assay. This is crucial to be able to effectively troubleshoot any issues. Finally, through this validation study, we realized that WGS often pushes microbiology to an unknown territory, in which many novel species were isolated in the clinical laboratory but yet there are no reference genomes available to match them, leaving an accurate ID unresolved. A consortium from the clinical microbiology laboratory community is needed to share these WGS data to help define and classify novel species.

To summarize, the implementation of NGS in a clinical microbiology setting is a challenging but ultimately necessary task. Lack of guidelines and standards, as well as infrastructure, technical, and financial barriers, contribute to the difficulty of implementation. This study provides an example of how to model validation studies for the implementation of bacterial ID by WGS in such a setting.

Acknowledgment

We thank the entire team of the UCLA Clinical Microbiology Laboratory NGS Division for making this WGS test available for clinical applications.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2021.07.020*.

References

1. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ: Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012, 8:e1002824
2. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW: Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 2012, 13:601−612

3. Caboche S, Audebert C, Hot D: High-throughput sequencing, a versatile weapon to support genome-based diagnosis in infectious diseases: applications to clinical bacteriology. Pathogens 2014, 3:258−279
4. Bartels MD, Petersen A, Worning P, Nielsen JB, Larner-Svensson H, Johansen HK, Andersen LP, Jarløv JO, Boye K, Larsen AR, Westh H: Comparing whole-genome sequencing with Sanger sequencing for spa typing of methicillin-resistant Staphylococcus aureus. J Clin Microbiol 2014, 52:4305−4308
5. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O: Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 2012, 50:1355−1361
6. McDonnell J, Dallman T, Atkin S, Turbitt DA, Connor TR, Grant KA, Thomson NR, Jenkins C: Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of Shigella sonnei in the UK. Epidemiol Infect 2013, 141:2568−2575
7. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S: Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med 2013, 10:e1001387
8. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk KC, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Grard G, Leroy E, Schneider BS, Fair JN, Martínez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J Jr, Miller S, Chiu CY: A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 2014, 24:1180−1192
9. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM: Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol 2014, 52:139−146
10. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M, Pallen MJ: A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 2013, 309:1502−1510
11. Rampini SK, Bloemberg GV, Keller PM, Büchler AC, Dollenmaier G, Speck RF, Böttger EC: Broad-range 16S rRNA gene polymerase chain reaction for diagnosis of culture-negative bacterial infections. Clin Infect Dis 2011, 53:1245−1251
12. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW: Predicting antimicrobial susceptibilities for Escherichia coli and klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother 2013, 68:2234−2244
13. 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 TE, Crook DW, Walker AS, Golubchik T: Prediction of staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J Clin Microbiol 2014, 52:1182−1191
14. Gygli SM, Keller PM, Ballif M, Blöchliger N, Hömke R, Reinhard M, Loiseau C, Ritter C, Sander P, Borrell S, Collantes Loo J, Avihingsanon A, Gnokoro J, Yotebieng M, Egger M, Gagneux S, Böttger EC: Whole-genome sequencing for drug resistance profile prediction in mycobacterium tuberculosis. Antimicrob Agents Chemother 2019, 63:e02175-18
15. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A: Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol 2014, 52:2365−2370
16. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, Tsai YC, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW; NISC Comparative Sequencing Program, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre JA: Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing enterobacteriaceae. Sci Transl Med 2014, 6:254ra126
17. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA: Whole-genome sequencing to identify transmission of mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet 2013, 381:1551−1560
18. Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Walker AS, Peto TE, Paul J, Llewelyn MJ: Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of staphylococcus aureus in an intensive care unit. Clin Infect Dis 2014, 58:609−618
19. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Baño J, Pascual A, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA: Global dissemination of a multidrug resistant Escherichia coli clone. Proc Natl Acad Sci U S A 2014, 111:5694−5699
20. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al: Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 2014, 345:1369−1372
21. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB, D'Arc S, Brazier J, Brown D, Coia JE, Douce G, Gerding D, Kim HJ, Koh TH, Kato H, Senoh M, Louie T, Michell S, Butt E, Peacock SJ, Brown NM, Riley T, Songer G, Wilcox M, Pirmohamed M, Kuijper E, Hawkey P, Wren BW, Dougan G, Parkhill J, Lawley TD: Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet 2013, 45:109−113
22. Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME: A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. BMJ Open 2013, 3:e002175
23. Fricke WF, Rasko DA: Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. Nat Rev Genet 2014, 15:49−55
24. Gargis AS, Kalman L, Lubin IM: Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J Clin Microbiol 2016, 54:2857−2865
25. Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, Rusu LI: WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare? Pathogens 2014, 3:437−458
26. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DG, Merker JD, Nagarajan R, Palicki LA, Robetorye RS, Schrijver I, Weck KE, Voelkerding KV: College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med 2015, 139:481−493
27. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS: Global Microbial Identifier initiative's Working Group 4 (GMI-WG4). Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infect Dis 2015, 15:174
28. Adekambi T, Colson P, Drancourt M: rpoB-based identification of nonpigmented and late-pigmenting rapidly growing mycobacteria. J Clin Microbiol 2003, 41:5699−5708