Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection

Robert Schlaberg, MD, MPH; Charles Y. Chiu, MD, PhD; Steve Miller, MD, PhD; Gary W. Procop, MD; George Weinstock, PhD; the Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology; the Microbiology Resource Committee of the College of American Pathologists

● **Context.**—Metagenomic sequencing can be used for detection of any pathogens using unbiased, shotgun next-generation sequencing (NGS), without the need for sequence-specific amplification. Proof-of-concept has been demonstrated in infectious disease outbreaks of unknown causes and in patients with suspected infections but negative results for conventional tests. Metagenomic NGS tests hold great promise to improve infectious disease diagnostics, especially in immunocompromised and critically ill patients.

Objective.—To discuss challenges and provide example solutions for validating metagenomic pathogen detection tests in clinical laboratories. A summary of current regulatory requirements, largely based on prior guidance for NGS testing in constitutional genetics and oncology, is provided.

Data Sources.—Examples from 2 separate validation studies are provided for steps from assay design, and validation of wet bench and bioinformatics protocols, to quality control and assurance.

Conclusions.—Although laboratory and data analysis workflows are still complex, metagenomic NGS tests for infectious diseases are increasingly being validated in clinical laboratories. Many parallels exist to NGS tests in other fields. Nevertheless, specimen preparation, rapidly evolving data analysis algorithms, and incomplete reference sequence databases are idiosyncratic to the field of microbiology and often overlooked.

(*Arch Pathol Lab Med.* 2017;141:776–786; doi: 10.5858/arpa.2016-0539-RA)

Accurate diagnosis of infections, such as pneumonia, meningitis/encephalitis, and sepsis, in hospitalized patients is an area of great clinical unmet need. More than 1 million patients are admitted for pneumonia alone in the United States per year, incurring ~$10 billion in associated costs.[1,2] For encephalitis, the annual disease burden in the United States exceeds 20 000 hospitalizations, with a high rate of long-term sequelae and mortality and $2 billion in costs to the health care system.[3] These syndromes can be caused by a wide array of pathogens with indistinguishable clinical presentations, often necessitating testing with a large panel of diagnostic assays in an attempt to establish the diagnosis. Despite these efforts, the etiology remains unknown in 20% to 60% of cases,[4–8] resulting in delayed or ineffective treatments, increased mortality, and excess health care costs.

By combining unbiased sequencing, rapid data analysis, and comprehensive reference databases, metagenomics can be applied for hypothesis-free, universal pathogen detection, promising to improve diagnostic yield for syndromic testing of these infections.[9–16] However, high costs, long sequencing times, and slow, unwieldy data analysis tools have long made it impractical to apply these methods routinely in diagnostic laboratories.[17,18] Recent workflow improvements have addressed many of these barriers, and metagenomic-based tests in many laboratories are in active development. However, metagenomic test validation for clinical use that fulfills regulatory requirements poses

unique challenges, because standard approaches used for conventional tests may not apply.

Here we present examples of unbiased metagenomic testing for pathogen detection from respiratory secretions and cerebrospinal fluid to highlight the challenges and pitfalls of this approach for diagnosis of respiratory and neurologic infections, respectively. A special emphasis will be placed on validation of laboratory protocols, data analysis algorithms, and reference databases, and the establishment of rigorous quality control (QC) metrics. For an overview of next-generation (or massively parallel) sequencing methods used in diagnostic microbiology and a glossary of commonly used terms,[19,20] current bioinformatics tools for metagenomic sequence data analysis,[10,17,21–24] and evolving regulatory guidance and requirements,[25–28] please refer to the referenced literature. The example data shown in this manuscript comprise part of much larger independent validation efforts that are ongoing at ARUP Laboratories (Salt Lake City, Utah), IDbyDNA Inc (Sunnyvale, California), and the University of California, San Francisco. Examples have been condensed for brevity and are intended to illustrate general concepts.

FAMILIARIZATION AND PLANNING

In addition to the issues discussed herein, laboratories should clearly define the intended clinical use and range of pathogen types detected and reported by the unbiased metagenomic test. Factors that need to be considered include turnaround times, specimen requirements, library preparation protocols, sequencing platform and depth, the need for automation, QC approaches, data analysis workflows, costs, clinical relevance of results, and many others. Laboratories should clearly define the results review and reporting workflow, because this will affect the choice of data analysis solutions. Given that many workflow components (eg, library preparation, sequencing chemistry, data analysis) are continually being developed and optimized, it is important to be familiar with all available options and anticipate updated protocols for validation in the future. It also may be useful to document plans for validation of workflow and bioinformatics modifications in advance and bank aliquots of well-characterized specimens and data sets for use during the initial validation period.

ASSAY DESIGN

Pathogens of Interest

Laboratories should define the scope of pathogens reported by the assay, which will depend on the intended use, specimen type, and patient characteristics. The choice of pathogens may affect the preferred nucleic acid extraction method, sequencing strategy (RNA and/or DNA), need for target enrichment or host nucleic acid depletion, sequencing depth, reference database design, and data analysis tools. Importantly, the assay will have to distinguish between pathogens and organisms that constitute normal flora or reflect environmental or laboratory reagent contamination, and will need to develop methods to accurately classify closely related organisms at various taxonomic levels (eg, genus, species, and/or subspecies). Laboratories will also need to develop mechanisms for the continued review of relevant literature and reference databases to update the list of reported pathogens.

Sequencing Strategy

Desired turnaround times, sequencing depth, throughput, and error tolerance of data analysis tools will influence the choice of optimal sequencing platforms, read lengths, paired-end versus single-end sequencing, and the number of controls and specimens to be sequenced per run. Generating shorter, single-end sequence reads can save time and money. However, this may reduce taxonomic resolution of sequence classification software because longer sequence reads are more readily classified to species or subspecies level.[17] The accuracy of classification algorithms and reference sequence databases can be predicted using simulated sequence reads of varying lengths and error profiles.

Data Analysis

Metagenomic next-generation sequencing (NGS) data sets are usually composed of mostly host-derived sequences and a minor, sometimes minute fraction of pathogen sequences, constituting a "needle-in-a-haystack" problem. A range of potential data analysis solutions exist, but most share an initial step of separating host from microbial sequences and a second step in which microbial sequences are assigned ("classified") to the deepest taxonomic level possible, ideally species or subspecies level. The first step is often computationally intensive because tens of millions of sequences need to be analyzed. The second step of taxonomic classification is confounded by incomplete reference sequence databases, taxonomic assignments that only partially correlate with sequence similarities, and the high inherent genetic diversity of many microbes, especially viruses.

Laboratories need to decide whether to use in-house–developed or commercial data analysis solutions, nucleic acid or protein sequences for classification, limited marker gene or complete genome databases, and local implementations or cloud-based solutions. Decisions will likely depend on intended use of the test, turnaround time requirements, evolving regulatory requirements, and availability of hardware infrastructure and bioinformatics support. The need for bioinformatics support extends beyond development and validation, and is also necessary for maintenance and updates. Major topics pertaining to database selection are also discussed below.

ASSAY VALIDATION

Wet Bench Work

Accuracy.—The broad scope of metagenomic testing poses challenges to determine accuracy. Availability of positive specimens and analysis costs limit the number of pathogens that can be included in validation studies. A combination of known positive specimens, pathogen-negative patient specimens spiked with whole-organism or purified nucleic acid, and simulated data (in silico validation) may be necessary. The entire workflow from specimen extraction to report generation can be validated with known positive and spiked patient specimens. Because unexpected pathogens may be detected in the process, it is typically important to consider strategies for orthogonal confirmation of such unexpected agents using well-characterized, independent methods. Whenever possible, clinically validated or US Food and Drug Administration (FDA)–approved tests should be used for this purpose.

Table 1. Summary of In Silico Validation Results, Accuracy[a]			
Total	**Taxa**	**Simulated Samples**	**Accuracy, %**
RNA viruses	73	230	100
DNA viruses	29	70	100
Bacteria	82	329	99.1
Fungi	1	6	100

[a] The table shows numbers of species and viral types (taxa), number of different reference sequences used to generate simulated samples representing the genetic intraspecies or intratype variability (simulated samples), and percent of simulated samples that were correctly identified as positive (accuracy).

Nevertheless, the traditional, individual target-specific approach to validation is not feasible for metagenomic sequencing tests because it is impossible to include all potential organism types and variants in validation studies. Similar challenges in molecular genetic pathology have led the College of American Pathologists to recommend methods-based or in silico proficiency testing[29,30] as an alternative. This approach is based on a combination of real sequencing data with in silico–generated or modified sequences for proficiency testing, which allows for more in-depth testing of algorithms and sequence databases, because a large diversity of specimens can be generated in silico and critical data analysis steps can be performed with fewer resources. In addition to proficiency testing of existing tests, a methods-based approach can also be applied to the initial validation of new tests.[30] Laboratories may choose a methods-based approach for validation, combining the advantages of traditional validation strategies with in-depth (in silico) validation using simulated samples. Critical workflow steps are defined, and their risk of generating an incorrect result is assessed. Using this risk-based approach, representative pathogen types and variants (eg, DNA viruses, RNA viruses, gram-positive bacteria, gram-negative bacteria, yeast, etc) may be selected for testing of the entire workflow. If available, residual patient specimens are preferred for validation, but spiking with whole-organism or purified nucleic acid may be necessary if such specimens are not available. For validation of uncommon pathogens or unusual strains, data sets generated in silico may be used to challenge classification algorithms and reference sequence databases more thoroughly.

Example 1: Respiratory Specimens.—A combination of the following strategies was used to determine accuracy:

1. Residual specimens positive for the most common and clinically relevant viruses
2. Spiking of test-negative patient specimens with cultured virus to supplement approach
3. Real sequencing data spiked with simulated pathogen reads for extensive in silico validation

Residual Specimens.—We had previously tested the accuracy of respiratory virus detection by unbiased metagenomics with more than 100 nasopharyngeal swab specimens.[11] Here, an additional 73 bronchoalveolar lavage and nasopharyngeal swab specimens with known results based on polymerase chain reaction (PCR) culture-based tests were tested. All specimens with discrepant results were also tested with independent, validated PCR-based tests; 42 specimens were positive by both predicate tests, of which 39 (92.9%) were positive for the same respiratory virus by metagenomics. In addition, previously undetected viruses

were identified in 6 specimens. These included human coronaviruses (n = 2), enterovirus (n = 1), human metapneumovirus (n = 1), human parainfluenza virus type 4 (HPIV-4; n = 1), human parechovirus (n = 1), respiratory syncytial virus (n = 1), and rhinoviruses (n = 3). Independent reverse transcriptase PCR (RT-PCR) tests were available for these viruses. However, for confirmation of unexpected results, independent methods may need to be set up or specimens be sent to outside clinical reference laboratories for additional testing.

Spiked Specimens.—When spiking residual patient specimens for accuracy studies, it is common practice to target organism loads at a given level above the sensitivity of the test (eg, 10-fold above the limit of detection [LOD]). Given the complexities of determining the LOD for a metagenomic test (see below), establishing analytic and clinically relevant LODs for all potential targeted organisms may not be practical. Alternative approaches include the use of representative, positive patient specimens of the relevant specimen type as a source of spiking studies. This ensures that spiked specimens contain concentrations similar to but below those in available test-positive specimens, and that a larger number of independent specimens can be used to evaluate background matrix effects. Alternatively, specimens demonstrated not to contain the microorganism under study can be spiked at designated concentrations and tested by quantitative PCR (qPCR) to target clinically relevant spiking levels.

Simulated Specimens.—Both the relevant specimen matrix and relevant pathogens at meaningful levels should be included for in silico validation. Accuracy was tested with 635 simulated samples for 185 taxa and ranged between 99.1% and 100% (Table 1).

Example 2: Cerebrospinal Fluid Specimens.—*Residual Specimens.*—Eighty-four cerebrospinal fluid (CSF) specimens that had previously tested positive for a single primary pathogen by conventional clinical laboratory testing (eg, culture, PCR, and/or serology), and 21 specimens that had previously tested negative were analyzed using the metagenomics test.[18] For the 21 specimens that had previously tested negative, available negative results corresponding to each of 5 different pathogen types (DNA virus, RNA virus, bacteria, fungi, and parasite) were included in the evaluation, for a total of 181 comparisons, 84 positive and 97 negative. Compared with conventional microbiologic testing, the metagenomic test showed 70.2% sensitivity and 92.8% specificity per specimen. All specimens with discordant results were also retested using an orthogonal, clinically validated method (eg, universal 16S bacterial PCR) if there was enough specimen available. For the 32 comparisons that were discordant (32 of 181; 17.7%), 14 of 25 previously positive specimens (56%) were found to be negative by discrepancy testing, attributed to target instability of the biobanked specimen or an initial false-positive clinical laboratory result. Inclusion of these data yielded a revised sensitivity of 84.3% and specificity of 93.7%.

Spiked Specimens.—In order to validate the performance of the metagenomic assay for organisms that were not present in available clinical CSF specimens, we performed spike-in experiments. *Neisseria meningitidis, Streptococcus agalactiae* (Group B Strep), and *Candida albicans* were obtained from the American Type Culture Collection (Manassas, Virginia) and subcultured. Synthetic CSF matrix (Golden West Biologicals, Temecula, California) was used as the diluent

Table 2. Some of the Variables That Influence Analytic Sensitivity				
	Dependent on...			
	Test Design	Pathogen	Specimen Type	May Vary by Specimen
Effectiveness of nucleic acid extraction	X	X	X	X
Pathogen genome size		X		
Efficiency of library preparation	X	X	X	X
Number of sequence reads	X			X
Availability of reference sequences		X		X
Sequence uniqueness		X		X
Accuracy of classification algorithms	X	X		X
Required confidence for identification	X	X		
Specimen composition			X	X

Abbreviation: X, variable influences analytic sensitivity.

to prepare a final concentration of approximately 100 colony-forming units (CFUs) per milliliter for each organism. Additional acid-fast bacilli organisms included *Mycobacterium fortuitum* and *Mycobacterium abscessus* isolates derived from infected patients. Acid-fast bacilli organisms were diluted to approximately 1000 CFU/mL. The metagenomic test achieved 100% sensitivity and specificity for detection for these spiked organisms.

Analytic Sensitivity

Sensitivity of pathogen detection depends on a series of variables: efficiency of nucleic acid extraction, pathogen genome size (at the same organism load, more reads are generated from longer genomes), robustness of library preparation, total number of sequence reads generated from a given specimen (more reads ≈ higher sensitivity), specimen composition and background reads, availability of appropriate reference sequences, sequence similarity with related organisms (confident differentiation of close relatives requires greater sequencing depth than identification of

unique sequences), accuracy of classification algorithms, and required confidence for pathogen identification. Many of these variables are known or can be influenced by careful test design (Table 2).

In addition, some variables are specific to a given specimen (Table 2). With unbiased metagenomics methods, the specimen composition (ie, cellularity and relative abundance of pathogens, other microorganisms, and patient cells) can affect analytic sensitivity. This is because "shotgun" metagenomic library preparation is sequence nonspecific, resulting in both host nucleic acid from the patient and microbial nucleic acid becoming part of the sequencing library. As a result, the pathogen load may result in different sequence coverage depending on the total nucleic acid yield (Figure 1).

Thus, analytic sensitivity can be determined for a given pathogen (strain) at a given sequencing depth and (1) an average specimen (by using a specimen pool for serial dilutions), (2) a representative control specimen (by using a contrived matrix; eg, a known quantity of cultured cells), or
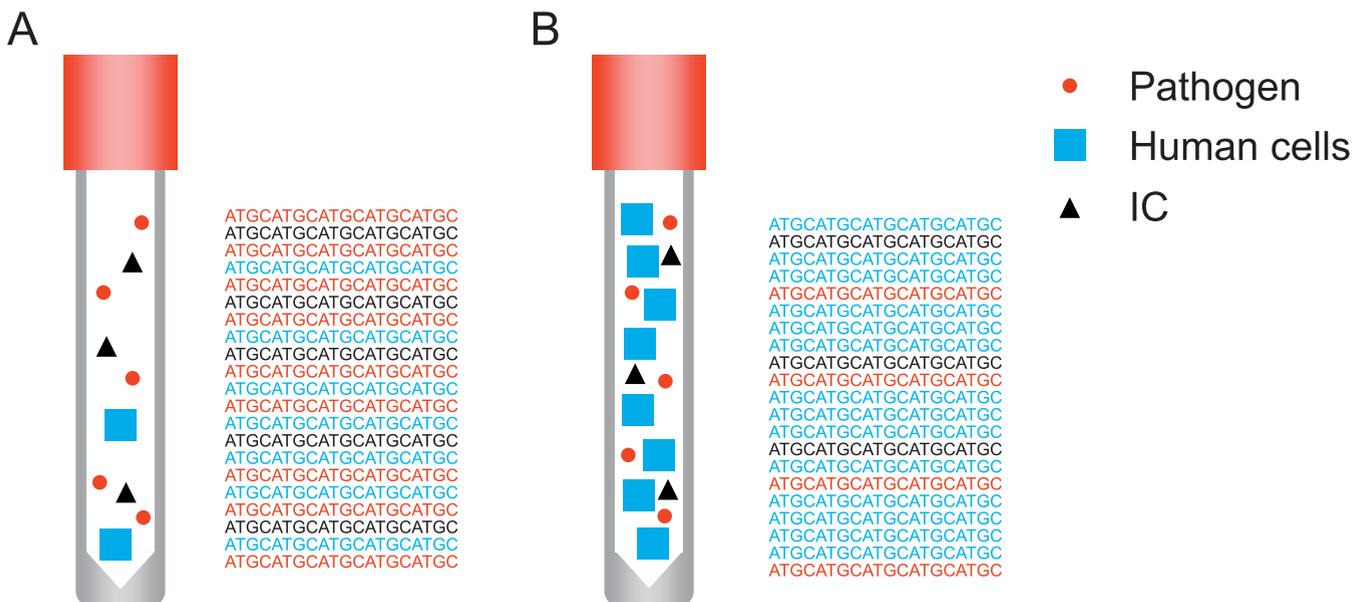

Figure 1. Conceptual illustration of how cellularity and overall nucleic acid yield affect analytic sensitivity in metagenomic next-generation sequencing tests. At the same pathogen load (red circles), specimens with fewer patient cells (blue) have higher analytic sensitivity (A) than specimens with a greater number of human cells (B). Because DNA/RNA from both pathogens and patient cells are sequenced, the number of pathogen sequences (red) will be lower and the number of human sequences (blue) will be be higher in samples with greater numbers of patient cells (B). Interal controls spiked into patient samples at known concentrations (IC, black) can be used to quantify this effect because the number of IC sequences (black) will also be reduced in samples with greater numbers of patient cells (B). During test validation, cutoffs can then be established for minimal numbers of IC sequences.

Table 3. Summary of In Silico Validation Results, Specificity[a]		
	Simulated Samples	**Specificity, %**
Viruses	312	100
Bacteria	1150	99.8
Fungi	409	99.8

[a] The table shows numbers of species and viral types not reported as respiratory pathogens by this test, which were used to generate simulated samples (simulated samples) and the percent of simulated samples that were correctly identified as negative (specificity). Rare false-positive results occurred because of misclassification of non-reportable pathogens with high sequence homology to reported respiratory pathogens. In silico validation results can be used to identify taxa that are prone to such false-positive results so that expert review can be focused on those results.

(3) an individual specimen or specimens (by using individual specimens for which sufficient amounts of residual volumes are available). As a result, determining limits of detection as routinely done for a PCR-based test may not be feasible.[31] Protocols are available to enrich pathogen nucleic acid,[32,33] deplete host nucleic acid,[34] or remove part of the sequencing libraries[35] to increase analytic sensitivity. However, none of these approaches completely eliminates the influence of specimen composition on analytic sensitivity. A pragmatic approach to overcome this limitation is the use of internal control(s) (IC) spiked at constant concentrations into all specimens (see QC below).

Example 1: Respiratory Specimens.—Analytic sensitivity was tested for 3 respiratory viruses (human metapneumovirus, HPIV-3, and respiratory syncytial virus), 3 bacteria (*Straphylococcus aureus*, *Klebsiella pneumoniae*, and *Haemophilus influenzae*), and 1 yeast (*Pneumocystis jirovecii*) by testing 10-fold serial dilutions prepared in (1) a pool of patient specimens and (2) a contrived matrix of cultured lung adenocarcinoma (A549) cells at a concentration that resulted in nucleic acid yields similar to average specimens. Results were compared to a validated qPCR test.

Example 2: CSF Specimens.—Limits of detection were determined using probit analysis for a mix of 7 representative pathogens, including cytomegalovirus (DNA virus), human immunodeficiency virus (HIV-1), *Streptococcus agalactiae* (gram-positive bacterium), *Klebsiella pneumonia* (gram-negative bacterium), *Cryptococcus neoformans* (yeast), *Aspergillus niger* (mold), and *Toxoplasma gondii* (parasite), that were spiked into negative CSF matrix (Golden West Biologicals Inc, Temecula, California) across approximately eight 5-fold serial dilutions (depending on the specimen), with an average of 3 replicates per dilution. These studies yielded overall LOD results for cytomegalovirus of 9.4 copies/mL; HIV-1, 100.8 copies/mL; *K pneumoniae*, 8.7 CFU/mL; *S agalactiae*, 8.9 CFU/mL; *C neoformans*, 0.01 CFU/mL; *A niger*, 130 CFU/mL; and *T gondii*, 55 organisms per milliliter.

Analytic Specificity

Limited analytic specificity may be a result of misclassification of microorganism or nucleic acid contained in patient specimens or contaminated reagents.[36–38] The former is largely a problem related to the classification algorithms used and/or reference databases available. Extensive in silico validation with a focus on microbes with sequence homology to relevant pathogens can help identify and mitigate problems. The latter concern can be addressed by including relevant negative controls and normalization metrics (see below). In addition, specificity challenges may be performed using patient specimens positive for or spiked with microorganisms that have sequence homology with relevant pathogens. If such organisms were already tested to demonstrate accuracy, results may also be used to demonstrate specificity. Performing an in-depth challenge to analytic specificity solely with spiked specimens is likely to be beyond the scope of most laboratories given the sheer number of spiked specimens that would need to be generated and tested.

Example 1: Respiratory Specimens.—Analytic specificity was demonstrated with patient specimens positive for relevant respiratory pathogens. No false-positive classification and identification of closely related pathogens (eg, human rhinovirus versus enteroviruses) was observed. In addition, a comprehensive in silico validation was performed using simulated specimens containing a wide range of viruses, bacteria, and fungi demonstrating 99.8% to 100% specificity (Table 3).

Example 2: CSF Specimens.—Specificity for speciation of related bacteria was determined using mixtures of *S aureus* and *Staphylococcus epidermidis* at 100%, 80%/20%, and 50%/50% (data not shown). The results demonstrated that the metagenomic test could reliably distinguish related species and identify coinfections, as long as the minority organism was present at a concentration of at least 20% relative to the predominant organism.

Reproducibility

As for other NGS tests, complex workflows pose challenges for reproducibility. Agreement between qualitative results (eg, detected, not detected) can be compared within and between runs. Laboratories may also want to assess reproducibility of quantitative variables used for results interpretation (eg, IC results). If quantitative results are reported, intrarun and interrun reproducibility as well as linearity should be assessed at multiple target concentrations.

Example 1: Respiratory Specimens.—Three patient specimens positive for representative pathogens were tested (from nucleic acid extraction) in triplicate within a single run and across 3 different runs. Results were compared qualitatively and were positive for all replicates (100% reproducibility). Because the analysis software uses quantitative variables to interpret results, provides normalized pathogen read counts, and uses quantitative results for IC interpretation, intrarun and interrun precision were determined for these measures.[17] Quantitative measures, such as depth and completeness of coverage, are algorithmically summarized into a single summary score ranging between 0 (no support for detection) and 1 (highest confidence for detection). Intrarun reproducibility was determined by testing the same patient specimens 3 times from nucleic acid extraction to sequencing on 1 run. As an example, the mean and % coefficient of variation of the summary score for detection of HPIV-4 in a positive patient sample were 0.6 and 25%, respectively. Interrun reproducibility was assessed with the same specimen, and mean and % coefficient of variation of the summary score for detection of HPIV-4 were 0.7 and 9%, respectively.

Example 2: CSF Specimens.—Positive control specimens consisting of 7 representative pathogens spiked as a mix into negative CSF matrix (the same organisms as used for the LOD studies) demonstrated 100% intrarun and interrun reproducibility for organism detection.

Validation of Metagenomic NGS Tests—Schlaberg et al

Table 4. Stability for Detection of Human Parainfluenza Virus Type 3			
Temperature	Time, d	Results	Summary Score
n/a	0	POS	0.7
4°C	4	POS	0.7
4°C	7	POS	0.7
−20°C	7	POS	0.8

Abbreviations: n/a, not applicable; POS, positive.

Stability

Specimen stability should be tested at time points and storage temperatures that are relevant given the clinical use of the test and logistical practicalities of specimen transport and storage. These studies can be performed with a limited number of representative pathogens. Known or suspected differences in stability may be used to select the least stable pathogens for stability testing.

Example 1: Respiratory Specimens.—In the example validation, 3 specimens were used to demonstrate stability for up to 7 days at refrigerated and frozen storage (Table 4 shows HPIV-3 results). Qualitative results agreed for all conditions, and the quantitative summary, a measure of classification confidence, showed no reduction over time.

Example 2: CSF Specimens.—Positive control specimens consisting of 7 representative organisms spiked into negative CSF matrix (the same organisms as used for the LOD studies) and their eluates were held refrigerated for up to 6 days or subject to 3 freeze-thaw cycles. All specimens demonstrated 100% detection of all 7 organisms under these conditions.

Bioinformatics

The following section provides a brief summary of important considerations for validating the accuracy and specificity of classification algorithms and reference databases. Relevant College of American Pathologists checklist requirements are listed in Supplemental Table 1, found in the supplemental digital content file (contains 2 tables; available at www.archivesofpathology.org in the June 2017 table of contents). Laboratories should consider validating the bioinformatics process (also termed "pipeline") using simulated specimens with known results and/or sequence data from previously analyzed clinical specimens prior to validating wet bench processes. The in silico validation takes advantage of the fact that next-generation sequence reads can be generated from reference sequences using publicly available simulation software (eg, wgsim[39]). Sequencing platform-specific error profiles (ie, frequency and types of sequencing errors) can be included for a more realistic challenge. Sequence reads may need to be simulated from reference sequences of the host (eg, human genome or transcriptome), commensal microorganisms if simulating nonsterile specimens, relevant pathogens, and closely related organisms (eg, when validating specificity). Another option to ensure realistic data sets is the combination of real sequencing data from pathogen-negative specimens with simulated sequence reads for pathogens of interest. Similar to wet bench spiking studies, the positivity (sequencing depth) can be controlled for simulated specimens. It may be advantageous to perform in silico validation studies prior to validation of wet bench processes, particularly when laboratories expect to modify databases or classification algorithms.

Covering all aspects of a thorough in silico validation is beyond the scope of this manuscript. However, because many challenges for classification algorithms and reference databases are specific to sequence-based microbial identification and differ substantially from applications in human genetics or oncology applications, important considerations are covered at least briefly here.

Accuracy

Reference databases for most pathogens are incomplete and biased.[40–42] Thus, many common pathogens may not have perfect matches in reference sequence databases. Generating simulated specimens solely based on reference sequence databases used for classification thus provides an unrealistic, simplified challenge. For more realistic in silico challenges, accuracy should take the expected genetic diversity of the target organism into consideration. Laboratories may also want to use independent classification tools or review alignment data using alternate methods to confirm correct annotation of pathogen sequences used for simulation studies, because reference sequences in public databases are frequently misannotated.[43] Simulated pathogen reads should be generated at coverages (ie, positivity levels) seen in patient specimens and reflect the length, paired or single-end characteristics, insert sizes, and error profiles of actual data used in the test. Lastly, simulated pathogen reads should be added to sequence reads that reflect the full background composition observed in patient specimens (see Example 1 below). In addition to simulated reads, reference data sets can be developed from metagenomic sequencing data corresponding to previously sequenced, well-annotated clinical specimens. These data sets with known results can then be used to benchmark any changes made in the bioinformatics algorithms or updates to the microbial reference databases.

Example 1: Respiratory Specimens.—In silico validation of accuracy included simulated pathogen reads combined with a "virtual specimen pool" to reflect sequences generated from patient cells, normal microbiota, and reagent contaminants that are part of "real" sequencing data (data not shown). The virtual specimen pool was prepared by combining 10 million nonpathogen reads randomly selected from 20 patient specimens. Simulated reads from relevant pathogens (n = 277) were then added into this virtual specimen pool to generate hundreds of simulated patient specimens. Pathogen reads were simulated at clinically relevant, low-positive levels. Simulated specimens were then analyzed and accuracy calculated by comparison to expected results.

Example 2: CSF Specimens.—Three reference data sets corresponding to metagenomic data from 24 well-characterized patient CSF specimens and 3 sets of positive and negative control replicates analyzed on 3 HiSeq runs (Illumina, San Diego, California) were developed as gold standard data sets to validate current and future changes to the bioinformatics analysis pipeline. Some examples of potential changes to the pipeline include: (1) regular updates to commonly used reference databases, such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/, accessed December 6, 2016); (2) incorporation of a confirmatory reference database standard, such as the FDA Database for Regulatory Grade Microbial Sequences (FDA-ARGOS; http://www.ncbi.nlm.nih.gov/bioproject/231221, accessed December 6, 2016); and (3) switching to a different
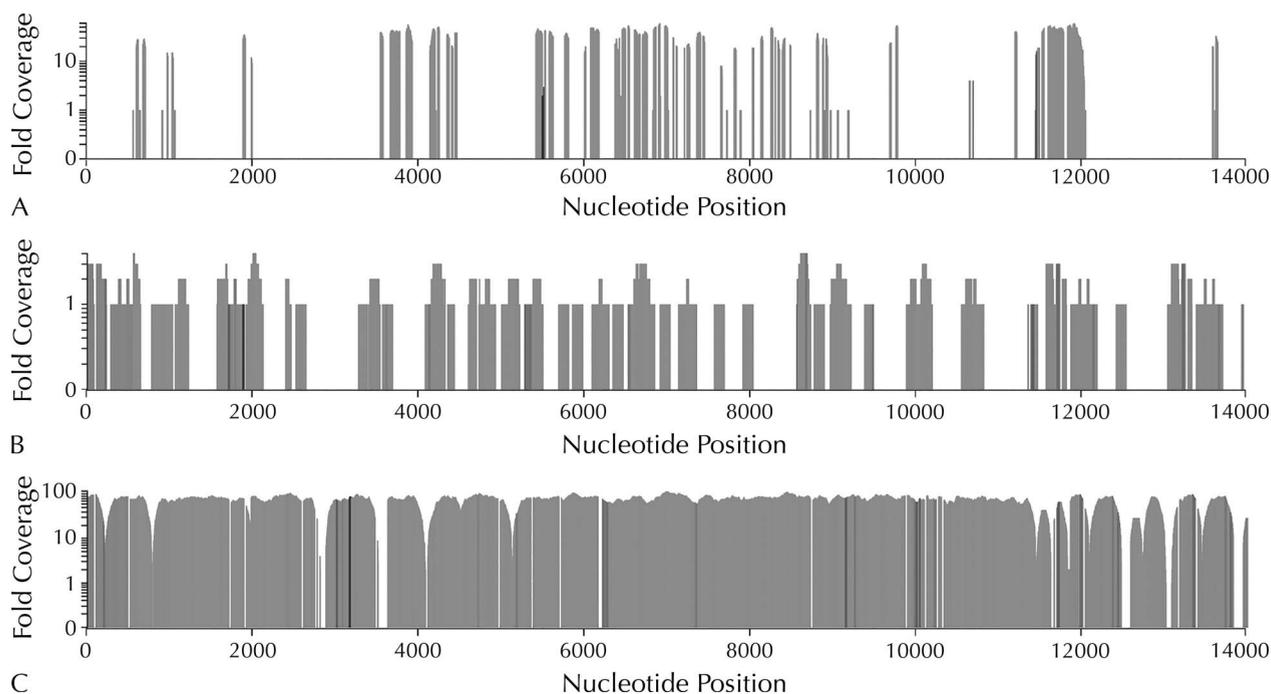
Figure 2. *Coverage plots of measles virus reference genome sequences when analyzing simulated samples positive for a false-positive algorithmic classification result due to a related morbillivirus with partial sequence homology (A; dolphin morbillivirus, simulated at 100× coverage across the entire genome), low-level true positivity (B; measles virus, simulated at 1× coverage across the entire genome) for measles virus, and high-level true positivity (C; measles virus, simulated at 10× coverage across the entire genome) for measles virus. The reference sequence for the measles virus genome is represented on the x-axis (nucleotide positions are shown). Coverage depth for simulated samples is shown on the y-axis. Differentiating false-positive detections based on partial sequence homology (A) from low-level, true-positive detections (B) is inherently challenging for classification algorithms. Results of a strong-positive sample (100× coverage depth) is shown for comparison (C). Expert review of coverage plots and alignments using defined criteria (standard operating procedure) may be necessary to maximize accuracy. High-level, true-positive detections can generally be made algorithmically with high confidence. Criteria and thresholds need to be validated, which can be most thoroughly done by using simulated specimens.*

nucleotide or translated nucleotide alignment algorithm for metagenomic analysis.

Specificity

Many microorganisms that could cause false-positive results will not be known a priori, may not have available reference sequences, or may be too numerous to test on the wet bench. In contrast, sequence reads can be simulated from hundreds or thousands of microbial genome sequences for more comprehensive in silico validation of specificity. It is critical to generate a comprehensive list of organisms with sequence homology or close taxonomic relationship that could cause false-positive results. Useful online resources to identify taxonomically related sequences and curated references include the Taxonomy Browser (http://www.ncbi.nlm.nih.gov/taxonomy; accessed December 6, 2016), SILVA,[44] the German Collection of Microorganisms and Cell Cultures (DSMZ; https://www.dsmz.de/, accessed December 6, 2016), the List of Prokaryotic Names with Standing in Nomenclature (LPSN; http://www.bacterio.net/, accessed December 6, 2016), StrainInfo (http://www.straininfo.net/, accessed December 6, 2016), the Viral Genome collection (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239, accessed December 6, 2016), ViralZone (http://viralzone.expasy.org/, accessed December 6, 2016), the Virus Pathogen Resource (https://www.viprbrc.org/, accessed December 6, 2016), and many others (see http://viralzone.expasy.org/all_by_species/677.html for examples, accessed December 6, 2016). To

challenge specificity, it is useful to simulate at high depth of coverage. Sequences may be spiked into a virtual specimen pool or tested by themselves. In addition, sequence reads from conserved genomic regions can often not unambiguously be classified to the species level. Therefore, it is important for classification algorithms to take the taxonomic context into account and classify sequence reads to the appropriate taxonomic depth.

Example 1: Respiratory Specimens.—Viruses that are closely related to the human respiratory viruses targeted by the metagenomics test were identified using the resources listed above. Sequence reads were generated in silico from full-length viral genome sequences (n = 312) simulating high-positive specimens (100× coverage) using the standard Illumina run error profile.[39] These studies were performed without a virtual specimen pool because (1) its effect had been tested as part of accuracy studies, (2) because the goal was to assess misclassification of high levels of organisms closely related to the pathogens of interest, and (3) because the data analysis tool used first groups host and microbial sequences prior to further classification.[17] Simulated reads from 7 animal viruses resulted in low-level positive detections using automated analysis criteria on the basis of sequence homologies. For this test, all results are manually reviewed, which led to identification of these false-positive detections per review criteria specified in the SOP (example in Figure 2 and Table 3). In addition, detection of most of these viruses from human specimens is highly unlikely.

Example 2: CSF Specimens.—Specificity of the metagenomic test is achieved by rapid taxonomic classification of sequence reads after alignment to a comprehensive reference database (http://www.ncbi.nlm.nih.gov/nucleotide/, accessed December 6, 2016) using a lowest common ancestor algorithm. Only reads that can be classified to the species and/or subspecies level are used for reporting a positive detection. Threshold requirements for clinical reporting of detected organisms are needed to maximize specificity. For viruses, reads spanning at least 3 nonoverlapping genomic regions must be identified. For bacteria, fungi, and parasites, a normalized ratio using a "no-template control" (NTC) sample processed in parallel is calculated to avoid false-positive detections due to reagent or laboratory contamination. First, the reads per million (RPM; the number of organism reads per million sequence reads) is determined for the sample and for the NTC. Then the RPM ratio is calculated as RPM_{sample}/RPM_{NTC}. A threshold value for the RPM ratio of 10 was determined empirically as the optimal cutoff for discriminating true-positive detections from background (S. Naccache et al, unpublished data, 2016). Manual expert review of assembled contigs (contiguous sequences) corresponding to informative genomic regions can be helpful, such as review of the assembled full-length enterovirus VP1 gene sequence for genotype-level identification.

Databases

An in-depth discussion of requirements, sources, strategies, and limitations of reference databases is beyond the scope of this manuscript. Some of the most relevant considerations are outlined briefly.

Database Selection

Numerous publicly available microbial reference sequence databases exist (see above). More comprehensive databases are preferred, but database size often comes at the cost of reduced annotation accuracy. In many databases, medically relevant and model organisms are vastly overrepresented, and genomic diversity of circulating strains is often not adequately represented. Skewed representation and diversity of reference databases can have a profound impact on the accuracy of many classification algorithms.[30] Strains not present in reference databases are often misclassified as closely related organisms that are represented in reference databases.[45] Thus, large, accurate, and diverse reference databases are desirable. The FDA's curated, public reference database (FDA-ARGOS, accessed December 6, 2016) represents an effort to generate highly curated organism sequence data, but it is currently limited in the scope of pathogens included (125 genomes, mostly bacterial, as of September 18, 2016).

Database Curation

Regardless of the source of reference sequences, laboratories may need to confirm the accuracy of annotations, completeness of sequences, comprehensiveness of reference databases, and representation of relevant taxa. Removal of misannotated species from the reference database or addition of clinically relevant organism sequences that are missing or underrepresented (eg, because of an emerging outbreak pathogen) may be used to generate a custom database, which needs to be periodically verified for accuracy and updated as needed. Clustering sequences based on user-defined sequence identity thresholds and using the resulting compressed database is one approach to limit redundancy and to decrease computational turnaround times. Depending on database size and source of reference sequences, automated or semiautomated methods may be required for curation.

QUALITY CONTROL

An NGS run will generally contain multiple, individually indexed (bar coded) sequencing libraries, consisting of 1 specimen with controls at a minimum. Reduced sequence quality can affect all specimens on a sequencing run (run-level) or be limited to individual specimens (specimen-level). Both run-level and specimen-level metrics for QC need to be established (Supplemental Table 2).

Sequencing Run-Level QC

QC of Run-Level Sequencing Quality and Yield.—Run-level QC is used to monitor the overall amount of sequence data generated (eg, total run yield, sequencing cluster density) and sequence quality (eg, % reads passing filter, % bases $\geq$Q30, sequencing error rate, sequencing read length). The utility of some metrics depends on the sequencing method (eg, read lengths vary more for some methods). In all cases, relevant recommendations and requirements of accrediting bodies need to be followed.[25–28]

Example: Respiratory Specimens.—Summary statistics were calculated for total run yields (>50 runs), cluster density, % reads passing filter, % bases $\geq$Q30, phiX error rate (a spiked specimen of genomic DNA from a reference phage routinely used to monitor sequencing error), and read lengths on 4 different NextSeq 500 instruments (Illumina) to identify outlier runs (Figure 3). Two runs were rejected based on excessive cluster density and less than 80% reads passing filter (Figure 3, highlighted). Cutoffs were derived based on mean $\pm$ 2 SDs and are comparable to manufacturer's specifications.[46] Cutoff for phiX error rate was set at 2.5%. As expected, read lengths were 150 bp for all runs.

External Controls.—External positive and negative controls should be included in each run. Laboratories may choose to rotate positive controls containing 1 or more pathogens. If quantitative values are used for results interpretation, acceptable ranges need to be established during the validation (eg, acceptable read counts for positive controls). Appropriate negative controls are critical to identify specimen-to-specimen and reagent contamination.[36–38,47] As discussed, analytic sensitivity (including sensitivity for detection of contaminants) depends on the specimen matrix and nucleic acid composition. Thus, negative controls need to simulate the relevant matrix, typically with low background to maximize sensitivity for detection of trace laboratory contaminants. Alternatively, extraction buffer or blank transport media may be used.

Example: CSF Specimens.—The variability in the number of sequences passing quality control metrics for positive and NTC samples over the course of 20 runs was analyzed. Based on these data, a threshold level of 5 million reads indicated sufficient metagenomic sequencing depth for organism detection. Negative samples with fewer than 5 million reads are reported with a comment indicating potential decreased assay sensitivity due to inadequate sequencing depth, and may be resequenced at higher depth at the discretion of the laboratory director after discussion with the clinical team.

Other QC Metrics.—Laboratories may also wish to monitor proportions of reads without a recognized bar

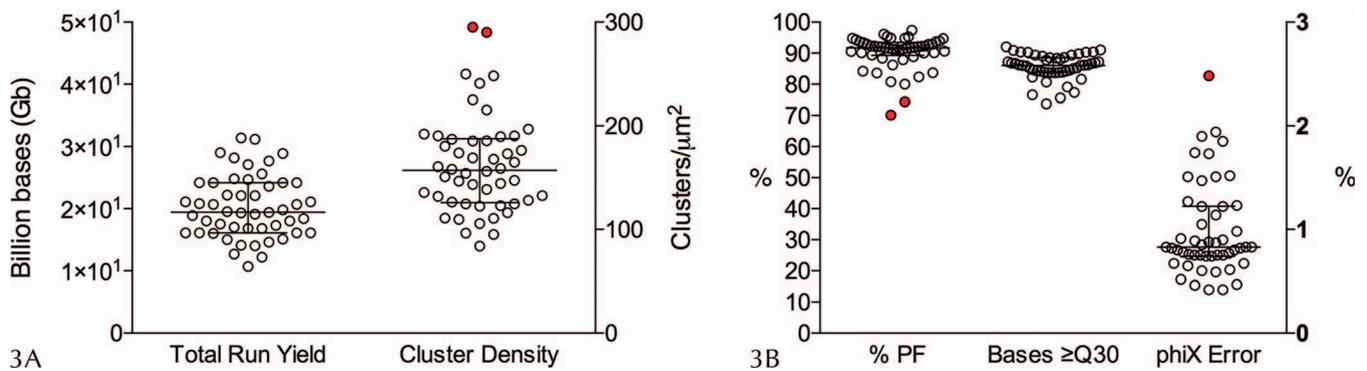*Validation of Metagenomic NGS Tests*—Schlaberg et al **783**

Figure 3. *Example of run-level quality control (QC) data; validation runs outside QC thresholds are highlighted. A, Distribution of mean run yields and cluster density for sequencing runs of bronchoalveolar lavage specimens processed for accuracy validation. B, Distribution of percent clusters passing filter, percent bases with Q scores equal to or greater than 30, and sequencing error rates as determined based on the phiX control are shown for sequencing runs that were part of the accuracy validation.*

code, which may indicate high sequencing errors, errors in bar code assignments, or specimen mixups.

Specimen-Level QC

QC of Extracted Nucleic Acid.—Depending on the specimen type, nucleic acid yields and quality may differ considerably between specimens. Methods for preparing NGS libraries from extracted nucleic acid generally require a minimum quantity and quality of nucleic acid. Acceptability criteria need to be defined in the validation.

QC of Sequencing Libraries.—Monitoring library yield and quality is required to guarantee adequate specimen pooling, ensure consistent quality, and identify problematic specimens or workflow errors. Library yield and quality can be assessed by direct methods (eg, Bioanalyzer, Agilent Technologies, Santa Clara, California; LabChip, Perkin-Elmer, Waltham, Massachusetts; or similar instruments) and/or by qPCR. Direct methods can provide information on the length of DNA fragments incorporated into sequencing libraries, which is dependent on the nucleic acid integrity of the original specimen and fragmentation protocols used during library preparation. This library size distribution can be used to identify low-quality specimens with degraded nucleic acid (ie, short library sizes), to identify incomplete nucleic acid fragmentation during library preparation (ie, long library sizes), and to normalize qPCR results, for which commercial kits are available. The SOP needs to specify library dilutions for quantification, numbers of replicates, concentrations and replicates for standards, and acceptability criteria.

QC of Specimen-Level Sequencing Data.—Because sequencing depth affects analytic sensitivity, it is important to validate the minimal number of required reads. Read counts for specimens analyzed on the same sequencing run are influenced by the total run yield (total number of sequence reads generated for all specimens; see above) and evenness of library pooling (ie, some specimens may have been overrepresented in the library pool, resulting in more sequence reads at the expense of other specimens on the same run). Pooling a greater number of specimens reduces costs but may also reduce evenness of pooling because of library quantification errors. When specimens fail minimal read count thresholds because of uneven pooling, read counts can be used to recalculate pooling ratios using the same libraries for repeat sequencing. Criteria and workflows should be developed for these situations. Depending on

data analysis strategies, proportions of unmapped or unclassified reads may be monitored because this can help identify the presence of unexpected organisms or high levels of contaminants. Read length distribution should also be monitored at the specimen level.

Internal Controls.—Spiking specimens with one or more ICs can help identify analytic failures and specimens with unusual cellularity that may result in reduced analytic sensitivity. The use of whole-organism controls spiked into the lysis buffer ensures that equal amounts of IC are added to all patient specimens and controls for the entire workflow, including nucleic acid extraction. Nucleic acid extraction efficiency may differ for relevant pathogens (eg, gram-positive bacteria may be harder to lyse), so laboratories may want to consider inclusion of representative organisms in the IC that account for differences in extraction efficiency. The amount of the IC added to specimens needs to be optimized to balance IC dropout (low spiking levels) and competition of IC with pathogens during unbiased sequencing (high spiking levels). Different IC mixes may be required for different sequencing approaches (eg, cDNA, DNA).

Example 1: Respiratory Specimens.—Sequencing libraries are assessed by LabChip (PerkinElmer) and qPCR. Median library size cutoffs were set to 250 to 600 bp based on a rounded mean ± 2 SDs (Figure 4, A). A minimum read count cutoff of 1.5×10^6 reads was chosen based on validation data (Figure 4, A), and accuracy and analytic sensitivity at low read counts. Internal controls consisting of a virus, gram-positive coccus, gram-negative rod, and yeast were spiked into the lysis buffer at fixed concentrations (Figure 4, B). Normalized read counts for IC organisms were determined by the analysis software and automatically interpreted according to predetermined threshold cutoffs.

Example 2: CSF Specimens.—An IC "spike-in" consisting of an RNA phage (M2) and a DNA phage (T1) was spiked into positive and negative control specimens and analyzed for signal variability. Results were analyzed for statistical variability and the ability of the assay to detect target organisms in the presence of potential interfering substances. Based on these parameters, a threshold of greater than 100 phage RPM was defined as sufficient detection of the IC to rule out assay inhibition. The DNA IC, in particular, was found to be a useful surrogate marker for human host background; for highly cellular CSF specimens (ie, high pleocytosis with white blood cell counts of >300/
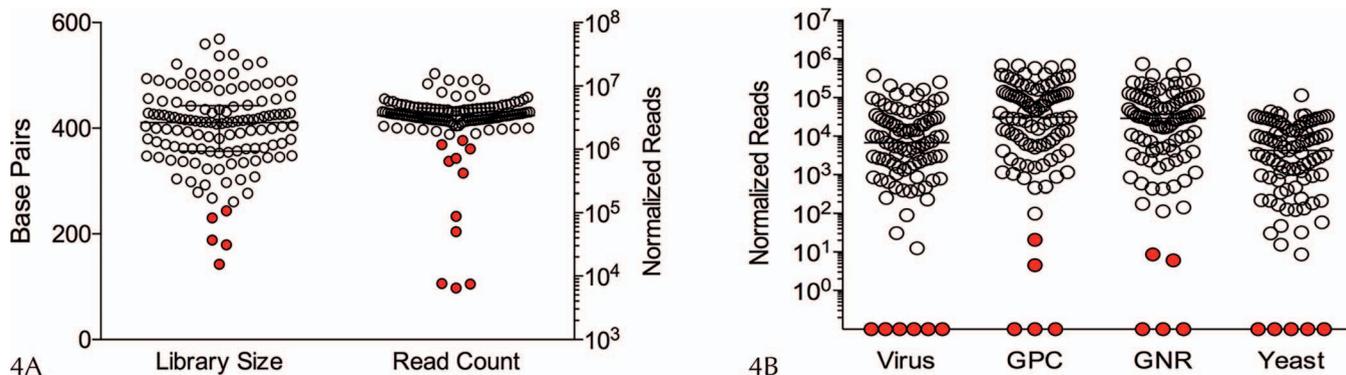
Figure 4. *Examples of specimen-level quality control (QC) results. A, Distribution of mean library sizes (left) and numbers of total reads per specimen (right) for bronchoalveolar lavagae specimens processed for accuracy validation. B, Distribution of normalized read counts for 4 internal controls used in RNA-seq analyses. Specimens highlighted in red fell outside a mean minus 2 SDs distribution and were repeated. Abbreviations: GNR, gram-negative rod; GPC, gram-positive coccus.*

mm³) the counts of T1 phage occasionally do not meet the passing criterion of greater than 100 RPM, and the sensitivity for detection of nonhuman reads from pathogens is reduced. The ICs are also used to monitor the effects of interfering substances, such as heme, from lysed red blood cells in CSF, which can inhibit the PCR reactions used during sequencing library preparation. Similar to the use of ICs in PCR assays, failure of the IC to meet detection criteria does not affect the reporting for samples with organisms detected, but a comment indicating decreased sensitivity for detection of that organism type is added to results that are negative for organisms belonging to that type.

RESULT REVIEW, REPORTING, PREPARATION FOR LAUNCH

The complex data generated by NGS-based pathogen detection tests need to be converted into interpretable and actionable information. This may require varying degrees of manual results review. Workflows for manual review, which include QC results, and reporting formats should be defined before starting validation studies. Requirements for and extent of manual result review will depend on test design, required turnaround times, functionalities of analysis software, laboratory infrastructure, and other factors. Workflow details, the exact roles of staff involved in results review, interpretative criteria, and requirements for confirmatory testing need to be specified in the SOP. If rapid turnaround times are planned, workflows that require less manual review and confirmation may be preferable. Processes and procedures for any unexpected results and review of their medical relevance may also need to be defined. Even extensively validated tests may result in identification of unexpected, nonvalidated microorganisms that may be of unclear medical relevance. Mechanisms and protocols for review and reporting, if appropriate, of these results should be defined and documented in the SOP.

OTHER CONSIDERATIONS

Other items to consider include criteria and detailed workflows for repeat testing in the event of QC failures. New reagent lots and shipments need to pass quality control prior to or in parallel with their first use. The SOPs and acceptability criteria for QC of reagents are required. Pragmatic approaches may be necessary given complex workflows involving numerous reagents and kits and their

high costs. However, the potential of lot-to-lot variability in quality and microbial nucleic acid contamination should be considered. One approach is to prioritize the extent of QC studies based on the risk that suboptimal reagent performance will have a major impact on assay performance. For example, new lots or shipments of enzyme-containing components may require more extensive QC than components that are less likely to fluctuate in quality and impact assay performance (eg, buffers). Laboratories may chose to perform QC on only high-risk components prior to first use, while accepting manufacturer QC data for low-risk components. It may also be beneficial to prepare multiple aliquots of informative real or contrived specimens with known results for reagent QC. The SOPs should contain a complete list of all reagents, kits, and equipment, and their sources.

When using in-house–developed data analysis algorithms and/or reference sequence databases, laboratories need to ensure that any changes are documented and that software and database versions are clearly identified and documented together with test results. In preparation for test launch, training and competency assessment of laboratory staff, including results review and reporting, needs to be planned and documented. In addition, sources of proficiency testing materials should be identified and proficiency testing challenges should be scheduled.

CONCLUSIONS

The potential of unbiased pathogen detection tests by metagenomic sequencing has been demonstrated in numerous studies and clinical contexts. Rapid reductions in costs and increases in throughput of NGS instruments, improved library preparation workflows, and massive increases in speed and ease of use of data analysis tools have brought these tests within the scope of clinical diagnostic laboratories. Although laboratory workflows, data analysis protocols, and regulatory requirements are still evolving rapidly, laboratories may want to apply the substantial promise of unbiased pathogen detection for testing of specimens from difficult-to-diagnose patients. The technology of NGS is rapidly becoming routine and shared between diagnostic specialties. In contrast, specimen preparation and data analysis steps are still highly complex, idiosyncratic to the field of microbiology, and often overlooked. It is the responsibility of diagnostic microbiol-

ogy laboratories to apply what has been learned in other areas of NGS testing to the unique challenges posed by highly variable specimen complexity and quality, broad genetic diversity of microorganisms, and imperfect and often highly incomplete reference databases. Clinical microbiologists are familiar with many of these challenges because they also apply to nucleic acid amplification tests and are in a unique position to apply their expertise to these emerging tests.

References

1. Pfuntner A, Wier LM, Stocks C. Most frequent conditions in U.S. hospitals, 2011. Rockville, MD: Agency for Healthcare Research and Quality; 2013. HCUP statistical brief no. 162.

2. Torio CM, Moore BJ. National inpatient hospital costs: the most expensive conditions by payer, 2013. Rockville, MD: Agency for Healthcare Research and Quality; 2016. HCUP statistical brief no. 204.

3. Vora NM, Holman RC, Mehal JM, Steiner CA, Blanton J, Sejvar J. Burden of encephalitis-associated hospitalizations in the United States, 1998-2010. *Neurology*. 2014;82(5):443–451.

4. van Gageldonk-Lafeber AB, Heijnen ML, Bartelds AI, Peters MF, van der Plas SM, Wilbrink B. A case-control study of acute respiratory tract infection in general practice patients in the Netherlands. *Clin Infect Dis*. 2005;41(4):490–497.

5. Louie JK, Hacker JK, Gonzales R, et al. Characterization of viral agents causing acute respiratory infection in a San Francisco University Medical Center Clinic during the influenza season. *Clin Infect Dis*. 2005;41(6):822–828.

6. Ewig S, Torres A, Angeles Marcos M, et al. Factors associated with unknown aetiology in patients with community-acquired pneumonia. *Eur Respir J*. 2002;20(5):1254–1262.

7. Glaser C, Bloch KC. Encephalitis: why we need to keep pushing the envelope. *Clin Infect Dis*. 2009;49(12):1848–1850.

8. Glaser CA, Honarmand S, Anderson LJ, et al. Beyond viruses: clinical profiles and etiologies associated with encephalitis. *Clin Infect Dis*. 2006;43(12):1565–1577.

9. Wilson MR, Naccache SN, Samayoa E, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med*. 2014;370(25):2408–2417.

10. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res*. 2014;24(7):1180–1192.

11. Graf EH, Simmon KE, Tardif KD, et al. Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: a systematic comparison to a commercial PCR panel. *J Clin Microbiol*. 2016;54(4):1000–1007.

12. Fischer N, Indenbirken D, Meyer T, et al. Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. *J Clin Microbiol*. 2015;53(7):2238–2250.

13. Phan TG, Luchsinger V, Avendano LF, Deng X, Delwart E. Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *J Gen Virol*. 2014;95(pt 4):922–927.

14. Alquezar-Planas DE, Mourier T, Bruhn CA, et al. Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing. *Sci Rep*. 2013;3:2468.

15. Greninger AL, Chen EC, Sittler T, et al. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One*. 2010;5(10):e13381.

16. Swaminathan S, Schlaberg R, Lewis J, Hanson KE, Couturier MR. Fatal zika virus infection with secondary nonsexual transmission. *N Engl J Med*. 2016;375(19):1907–1909.

17. Flygare S, Simmon K, Miller C, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol*. 2016;17(1):111.

18. Mongkolrattanothai K, Naccache SN, Bender J, et al. Neurobrucellosis: unexpected answer from metagenomic next-generation sequencing [published online ahead of print January 6, 2017]. *Pediatr Infect Dis J*. doi:10.1093/jpids/piw066.

19. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio*. 2015;6(6):e01888–15.

20. Westblade LF, van Belkum A, Grundhoff A, et al. Role of clinicogenomics in infectious disease diagnostics and public health microbiology. *J Clin Microbiol*. 2016;54(7):1686–1693.

21. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3):R46.

22. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*. 2015;16:236.

23. Huson DH, Beier S, Flade I, et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*. 2016;12(6):e1004957.

24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59–60.

25. New York State Department of Health. Validation of next generation sequencing (NGS)-based methods for identification and/or characterization of infectious agents (isolates only). 2016. http://www.wadsworth.org/sites/default/files/WebDoc/ID WGS NGS Molecular Guidelines for Isolates_0.pdf. Accessed December 6, 2016.

26. US Food and Drug Administration. Infectious disease next generation sequencing based diagnostic devices: microbial identification and detection of antimicrobial resistance and virulence markers; draft guidance for industry and Food and Drug Administration staff. 2016. http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM500441.pdf. Accessed December 6, 2016.

27. Aziz N, Zhao Q, Bry L, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med*. 2015;139(4):481–493.

28. College of American Pathologists (CAP). CAP accreditation checklists—2016 edition. http://www.cap.org/ShowProperty?nodePath=/UCMCon/Contribution Folders/WebContent/pdf/accreditation-checklists-full-listing.pdf. Accessed August 17, 2016.

29. Duncavage EJ, Abel HJ, Merker JD, et al. A model study of in silico proficiency testing for clinical next-generation sequencing. *Arch Pathol Lab Med*. 2016;140(10):1085–1091.

30. Schrijver I, Aziz N, Jennings LJ, Richards CS, Voelkerding KV, Weck KE. Methods-based proficiency testing in molecular genetic pathology. *J Mol Diagn*. 2014;16(3):283–287.

31. Clinical & Laboratory Standards Institute. *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline–Second Edition*. Wayne, PA: Clinical & Laboratory Standards Institute; 2012. Publication no. EP17-A2.

32. Briese T, Kapoor A, Mishra N, et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio*. 2015;6(5):e01491–15.

33. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing using targeted sequence capture. *Genome Res*. 2015;25(12):1910–1920.

34. Hasan MR, Rawat A, Tang P, et al. Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J Clin Microbiol*. 2016;54(4):919–927.

35. Gu W, Crawford ED, O'Donovan BD, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol*. 2016;17:41.

36. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*. 2014;9(10):e110808.

37. Strong MJ, Xu G, Morici L, et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog*. 2014;10(11):e1004437.

38. Naccache SN, Greninger AL, Lee D, et al. The perils of pathogen discovery: origin of a novel parvovirus-like genome traced to nucleic acid extraction spin columns. *J Virol*. 2013;87(22):11966–11977.

39. Heng Li. Wgsim Reads simulator. https://github.com/lh3/wgsim. Accessed December 6, 2016.

40. Kyrpides NC, Hugenholtz P, Eisen JA, et al. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol*. 2014;12(8):e1001920.

41. Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 2013;499(7459):431–437.

42. Anthony SJ, Epstein JH, Murray KA, et al. A strategy to estimate unknown viral diversity in mammals. *MBio*. 2013;4(5):e00598–13.

43. Petti CA, Bosshard PP, Brandt ME. *Interpretive Criteria for Identification of Bacteria and Fungi by DNA Target Sequencing: Approved Guideline*. Wayne, PA: Clinical and Laboratory Standards Institute; 2008:viii.

44. Yilmaz P, Parfrey LW, Yarza P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*. 2014;42(database issue):D643–D648.

45. Ackelsberg J, Rakeman J, Hughes S, et al. Lack of evidence for plague or anthrax on the New York City subway. *Cell Syst*. 2015;1(1):4–5.

46. illumina. NextSeq Series of Sequencing Systems. http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-nextseq-500.pdf. Accessed December 6, 2016.

47. Rosseel T, Pardon B, De Clercq K, Ozhelvaci O, Van Borm S. False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. *Transbound Emerg Dis*. 2014;61(4):293–299.