AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

| Biotechnology | Research Article

Analytical assessment of metagenomic workflows for pathogen detection with NIST RM 8376 and two sample matrices

Jason G. Kralj,[1] Stephanie L. Servetas,[1] Samuel P. Forry,[1] Monique E. Hunter,[1] Jennifer N. Dootz,[1] Scott A. Jackson[1]

AUTHOR AFFILIATION See affiliation list on p. 13.

ABSTRACT We assessed the analytical performance of metagenomic workflows using NIST Reference Material (RM) 8376 DNA from bacterial pathogens spiked into two simulated clinical samples: cerebrospinal fluid (CSF) and stool. Sequencing and taxonomic classification were used to generate signals for each sample and taxa of interest and to estimate the limit of detection (LOD), the linearity of response, and linear dynamic range. We found that the LODs for taxa spiked into CSF ranged from approximately 100 to 300 copy/mL, with a linearity of 0.96 to 0.99. For stool, the LODs ranged from 10 to 221 kcopy/mL, with a linearity of 0.99 to 1.01. Furthermore, discriminating different *E. coli* strains proved to be workflow-dependent as only one classifier:database combination of the three tested showed the ability to differentiate the two pathogenic and commensal strains. Surprisingly, when we compared the linear response of the same taxa in the two different sample types, we found those functions to be the same, despite large differences in LODs. This suggests that the "agnostic diagnostic" theory for metagenomics (i.e., any organism can be identified because DNA is the measurand) may apply to different target organisms and different sample types. Because we are using RMs, we were able to generate quantitative analytical performance metrics for each workflow and sample set, enabling relatively rapid workflow screening before employing clinical samples. This makes these RMs a useful tool that will generate data needed to support the translation of metagenomics into regulated use.

IMPORTANCE Assessing the analytical performance of metagenomic workflows, especially when developing clinical diagnostics, is foundational for ensuring that the measurements underlying a diagnosis are supported by rigorous characterization. To facilitate the translation of metagenomics into clinical practice, workflows must be tested using control samples designed to probe the analytical limitations (e.g., limit of detection). Spike-ins allow developers to generate fit-for-purpose control samples for initial workflow assessments and inform decisions about further development. However, clinical sample types include a wide range of compositions and concentrations, each presenting different detection challenges. In this work, we demonstrate how spike-ins elucidate workflow performance in two highly dissimilar sample types (stool and CSF), and we provide evidence that detection of individual organisms is unaffected by background sample composition, making detection sample-agnostic within a workflow. These demonstrations and performance insights will facilitate the translation of the technology to the clinic.

KEYWORDS metagenomics, reference material, spike-in, pathogen detection

Next-generation sequencing-based metagenomics (also known as metagenomic sequencing, mNGS) pathogen detection offers tremendous promise as an "agnostic diagnostic," enabling relatively unbiased sample characterization (1, 2). This approach typically works by extracting samples' DNA and/or RNA, preparing them for a sequencer

Editor Ana Cabrera, London Health Sciences Centre, London, Ontario, Canada

Address correspondence to Jason G. Kralj, Jason.kralj@nist.gov.

NIST produced and sells RM 8376 on a cost-recovery basis.

Received 10 December 2024
Accepted 20 January 2025
Published 10 March 2025

(i.e., library preparation), collecting the raw sequence data for the sample, and then comparing the sequences to databases to assign likely identities of the sample constituents. This contrasts with *targeted* sequencing, which is effectively a highly multiplexed PCR assay that requires *a priori* knowledge of the targets. Data can be reanalyzed using multiple bioinformatics tools and/or databases, making comparisons and refinement of workflows possible without rerunning samples.

To aid the translation of metagenomic workflows into regulated applications, there is a clear need for methods assessing the analytical performance of a workflow using appropriate controls—organisms of interest, at appropriate levels, and in matrix. And though there are limited quantitative reference materials for mNGS applications, the release of NIST Reference Material (RM) 8376 aimed to remedy that (3).

RM 8376 consists of 19 individual tubes of pathogenic bacterial DNA extracted from cultured isolates and one tube of human DNA, with quantified genome copy number concentrations. As part of its mission to promote U.S. innovation and industrial competitiveness, NIST has been working with stakeholders in industry, academia, and other government agencies to develop standards to support the advancement of mNGS for clinical diagnostics and biosurveillance applications. This material can be employed to assess taxon-specific performance and identify potential weaknesses in a workflow before moving into clinical evaluation (4).

RMs, used as spike-ins or internal controls, have begun to find utility to assess the performance in metagenomics of mNGS tools (5, 6). In these studies, the response to spike-in concentrations in mock samples was assessed for 16S-based sequencing in performing absolute quantitation on samples. This marked a significant improvement to enable comparison studies and appears poised to enable the reproducibility of metagenomics analyses.

These RMs support workflow development by both facilitating the (i) assessment of analytical performance and (ii) identification of analytical biases in the mNGS workflow (7–11). It is important to establish analytical performance for each mNGS workflow due to the biases that result from each step (e.g., sample collection/handling, extraction, processing, and the bioinformatics). Together, these biases cause a discrepancy between the reported and actual sample compositions, ultimately impacting workflow performance. The impact of workflow choices, such as library preparation and sequencing depth, can be assessed and optimized using these RMs before analyzing clinical samples. With respect to the analytical performance, the first steps include establishing the LOD, calibration curve, and dynamic range; without establishing analytical performance characteristics, translation into a clinical setting becomes extraordinarily challenging.

Different sample types can have dramatically different profiles, with a wide range of organism concentrations and diversity (12–15), resulting in different biases to overcome (Fig. 1).

For applications involving sterile fluids, the confirmed presence of any microbe would be a cause for concern. In blood and CSF where samples should be nearly sterile, any DNA signal from a suspected pathogen above background is a cause for concern (9, 12, 16, 17), making LOD a critical parameter. This makes CSF an ideal candidate for diagnostics based on metagenomics, as the background signal is low, thereby simplifying the review of post-metagenomics analysis.

For samples with diverse microbial populations, only the presence of specific pathogenic strains would necessitate action. Such is the case in medium- to low-complexity samples (10 s of taxa) such as urine or vaginal swabs. Here, having specificity for the suspected pathogen is a critical performance parameter, with LOD supporting the presence/absence, especially if attempting to identify antimicrobial-resistant (AMR) strains (13).

In high-complexity samples such as stool (100s to 1,000s of species) (15), the analytical methods can dramatically impact the analysis (6–8, 18). Quantifying a critical abundance of specific strains such as shiga-toxin producers, viruses, and parasites may be a higher priority than having a low LOD, as with previous scenarios. Pathogenic strains

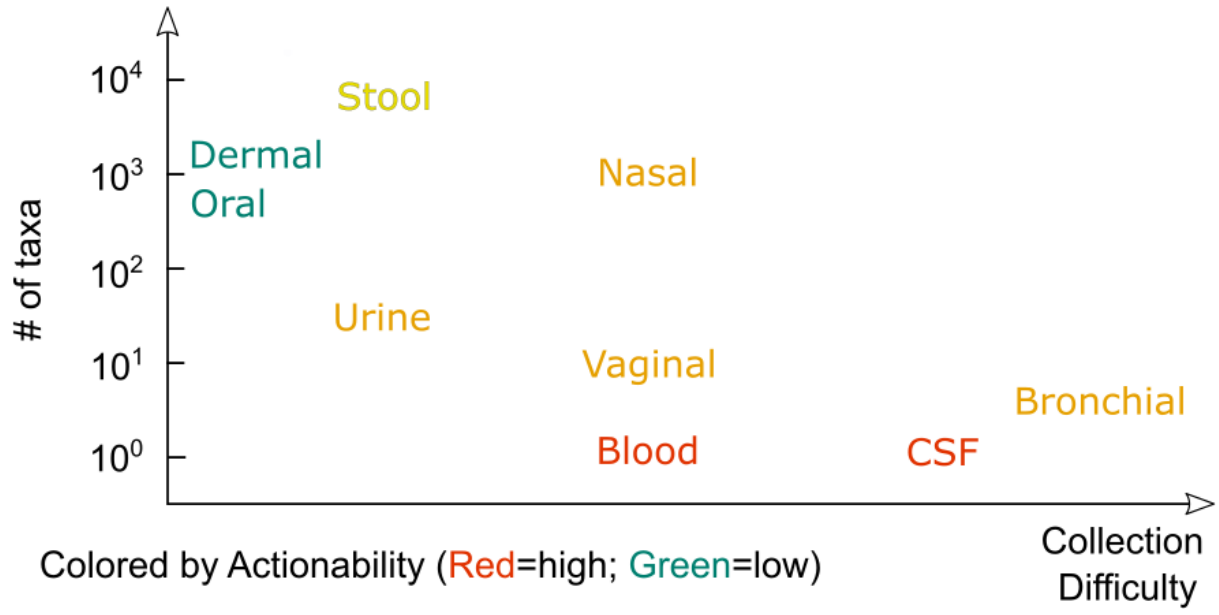


FIG 1 A subjective view of where human specimens investigated by mNGS typically fall in terms of complexity, difficulty of collection, and actionability. The number of taxa for a given sample type is approximate and may change as our understanding of each microbiome evolves. The actionability gives an approximate "score" for our understanding of the likelihood of finding pathogens at that site with potential for harm; individual use cases will vary, and changes in technology will impact each factor.

of many enteric pathogens may differ only slightly in their genomes from otherwise commensal flora, making identification challenging. Therefore, basic analytical performance assessment, with some nuances specific for each application, remains a consistent need in all mNGS applications. These matrices (CSF and stool) represent extremes in terms of microbial abundance and complexity and served to bracket the range of clinical sample types described earlier (Fig. 1).

Individual laboratory assessments have previously shown that purpose-built workflows (i.e., fit-for-purpose for specific sample types) perform well on clinical sample types and have been approved for diagnostic use through CAP-CLIA. These laboratories have targeted clinical needs for detecting nonhuman DNA in otherwise sterile body fluids, specialized in given sample types, and have performed rigorous workflow characterization and validation (12, 17, 19). However, to date, there are no examples of Food and Drug Administration (FDA)-cleared testing devices, highlighting the difficulty in translating this diagnostic for broad use.

In this work, we demonstrated how components of RM 8376 can be used to assess the analytical performance of an mNGS workflow: limits of detection, response curve, and linear dynamic range in two different sample types, cerebral spinal fluid (CSF) and stool. While only bacterial DNA was used in this work, the same process could be applied to other fungi, viruses, and parasites using an appropriate set of control materials and experimental workflow. The strategy utilized RM 8376 as both internal controls (*A. hydrophila* and *L. pneumophila*) and variable spike-ins (*K. pneumoniae*, *E. faecalis*, *N. meningitidis*, and *L. monocytogenes*) in two clinically relevant matrices: CSF and stool. Since most of these organisms are not commonly found in CSF and stool, we could decouple each organism's signal from background. The main goal of this work was to demonstrate the utility of genomic reference materials and potential experimental designs; however, we also observed some interesting, and in some cases, unexpected results about common classifiers—namely, that the analytical response for each taxa was the same regardless of the matrix despite LODs differing by >100 fold.

RESULTS AND DISCUSSION

Experimental overview

Cerebrospinal fluid (CSF) use case

The determination of the LOD for each taxon consisted of three steps:

1. Estimate the background signal for each taxon and determine the minimum signal for detection/quantitation.
2. Perform a linear regression on the data above the minimum signal.
3. Use the linear model with the minimum signal to determine the LOD/LOQ.

We examined the background signal for all samples. (Fig. 2, top) We observed the samples with high spike-in concentrations, which appeared to generate background normalized abundances much greater than low spike-in samples. We determined it was only appropriate to include samples with spike-in abundances less than 10^4 copy/µL because these samples' compositions were nearly the same and would more closely resemble a "typical" clinical sample (see SI). Typically, 0 or one absolute reads were observed for each taxon in any of these background samples, indicating the CSF samples were typically free of these species. From this, the minimum signal above background for each taxon was calculated. We note that the analytical performance metrics found in this use case apply only valid for each specific workflow. Significantly increasing the sequencing read depth by tenfold or even 100-fold would likely improve the LOD here because so few absolute reads were observed in the background, though the per-sample cost may significantly increase as well.

A regression of the data above the minimum signal was then calculated and plotted. (Fig. 2, bottom). For these four taxa, the slope of the response was approximately 1.0 with an uncertainty of approximately 1% to 5% (see Fig. 2, bottom), indicating no saturation of signal over the range of spike-in concentrations tested.

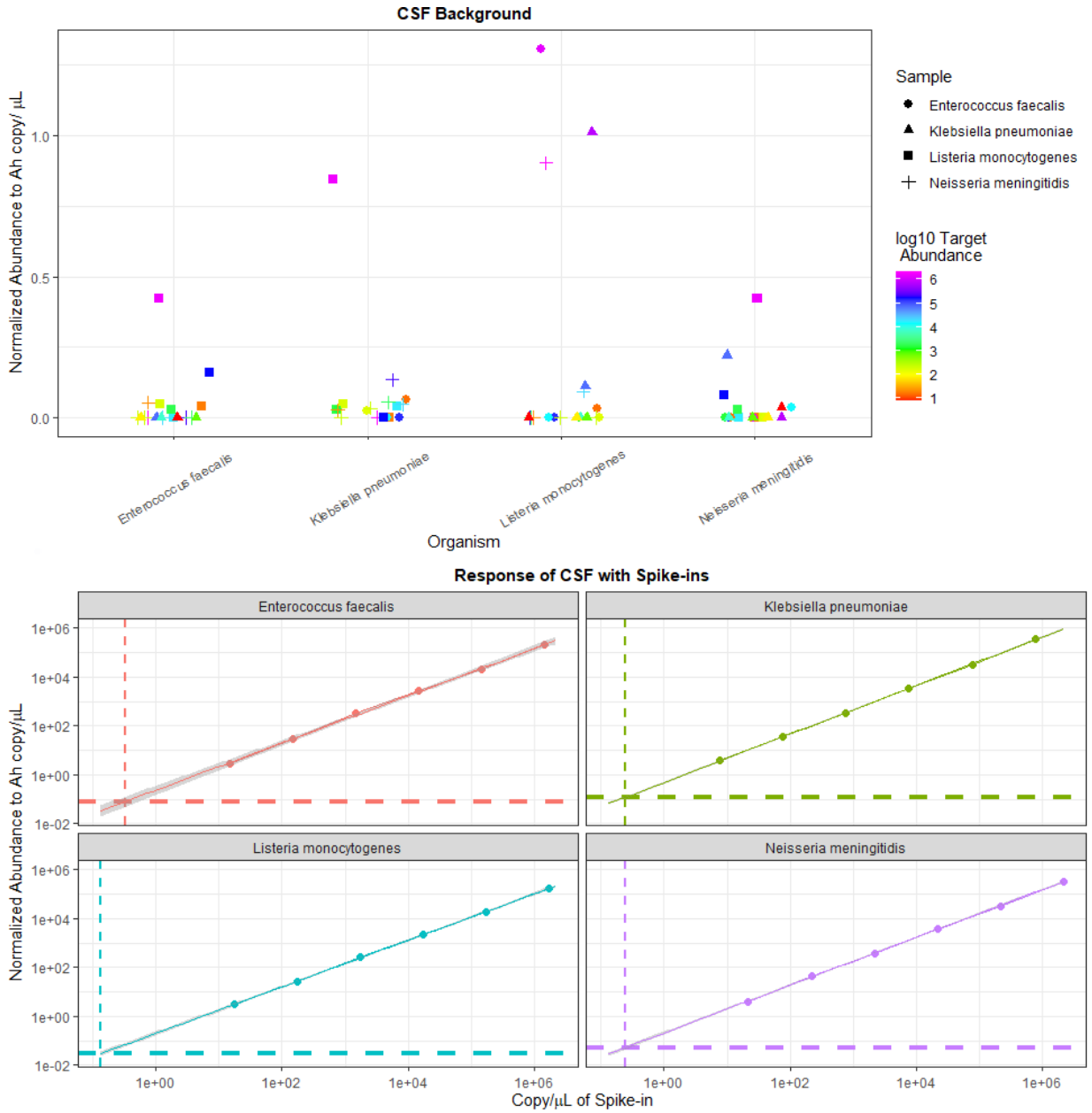
The linear model was used with the minimum signal to estimate the LOD for each taxon (Fig. 2, bottom). The LODs ranged from approximately 0.1 to 0.3 copy/µL. These values were statistically different from each other, though of the same order of magnitude.

The linear dynamic range would typically also be assessed for where the response signal saturates or becomes unstable; however, we were unable to reach this maximum value. Since the typical operating range for this matrix is at or near the LOD, which is approximately six orders of magnitude below the maximum value we tested, we concluded that LDR was sufficient for low-complexity CSF samples.

The same data can be used to compare alternative workflows (i.e., quality filtering, subsampling, classifiers, and databases) with respect to answering the question of which workflow is better suited for a given sample set. We assessed the performance using the classifier Kraken2 (Fig. S1) and compared the results to the workflow using Centrifuge. The results for CSF data were nearly identical both in terms of LOD and linearity for the four taxa studied. Hence, either/both data workflows would be comparable in an investigation involving CSF with those taxa.

It is worth noting that this analysis accounts for the performance of the sequencing and bioinformatics. Extraction efficiency is a known variable that impacts the response, and it may be necessary to increase the read depth of each sample to account for this in clinical samples. In this use case, each taxon appears to respond similarly. We would expect the response of most other bacterial DNA to be similar in CSF using the specific sequencer and experimental methods used. However, the first step here was to assess the LOD and response using a quantified DNA standard—and to that end, the experiment was successful in assessing these analytical performance metrics.

While we do not seek to establish this workflow as a diagnostic, evaluating the overall analytical performance in terms of comparability to existing workflows highlights the utility of reference materials for assessing the comparability of workflows. And given the



Organism	Bkgd normAb to Ah mean (sd) copy/µL	Linear Regression m (CI)	Linear Regression b (CI) copy/µL	LOD mean (CI) copy/µL
E. faecalis	0.014 (0.022)	0.97 (0.045)	0.24 (0.069)	0.32 (0.09)
K. pneumoniae	0.036 (0.029)	0.99 (0.010)	0.49 (0.062)	0.25 (0.03)
L. monocytogenes	0.003 (0.009)	0.95 (0.019)	0.20 (0.023)	0.13 (0.01)
N. meningitidis	0.008 (0.015)	0.98 (0.014)	0.21 (0.018)	0.24 (0.02)

FIG 2 Individual organisms were spiked into CSF at different concentrations ranging from approximately 10^6 to 10^1 copy/µL. (Top) The background signal for each taxon was plotted as the normalized abundance to *Aeromonas hydrophila* (Ah). Control samples spiked with high amounts of other taxa (i.e., not the target taxa) generated significantly higher background signals compared with the control samples with low levels of spiked amounts. Because of this, we identified

(Continued on next page)

Fig 2 (Continued)

two criteria for including data in the LOD calculation—the target taxon could not be spiked into that sample, and the spiked-taxon concentration was below 10^4 copy/µL (excluding blue and purple data points) so that the sample profile approximately resembles clinical samples with low levels of pathogens. (Middle) The response for each spiked-in taxon was plotted as normalized abundance to Ah vs the spiked-in copy number. A linear regression with confidence band was included. The filtered background signal (horizontal dashed line) and the linear regression of the data were used to estimate the LOD for each organism (vertical dashed line). The LODs were similar for each organism, ranging from approximately 100 to 300 copy/mL. Subtle differences between the performance of each taxon were likely significant and due to how unique each read could be interpreted by the classifier, although this effect was small. (Bottom) Analytical performance metrics for the Centrifuge classifier for CSF. The background signal relative to Ah, linear regression (b + m·log x), and LOD for the four taxa are shown here. The values are shown as mean (95% confidence interval), unless otherwise noted. The slopes of the regression (m) showed a response of approximately 1 for the range of spike-in concentrations studied. The LODs correspond to signals with fewer than 10 total reads, indicating that increasing the number of reads per sample would likely lower the LOD; however, this would require additional study, increased cost, and may not be necessary based on the particular criteria for a given use case.

simplicity of using the quantified spike-ins, assessing the preclinical analytical performance of the metagenomics portion of the workflow would inform on potential challenges (i.e., background signal and need for pre-amplification) before using patient samples.

One similar workflow from a previous clinical study estimated their LOD in CSF as between 0.1 and 0.2 genome copy/mL cell equivalents based upon 10 RPM-r (reads per million-relative to baseline), read length, and an average bacterial genome size of 4 Mbase (17). However, two rounds of universal PCR were used to pre-amplify the targets, which we did not employ here, resulting in the 1,000-fold lower detection threshold. Focusing on the 10 RPM-r LOD figure from the sequencing data alone, we achieved a similar analytical sensitivity (approximately 3 RPM, based upon the baseline estimate for minimum signal, LOD, and measured reads from samples) with this workflow. Hence, when accounting for the differences in sample preparation, the analytical performance derived from the contrived samples used here compared well with the clinical studies.

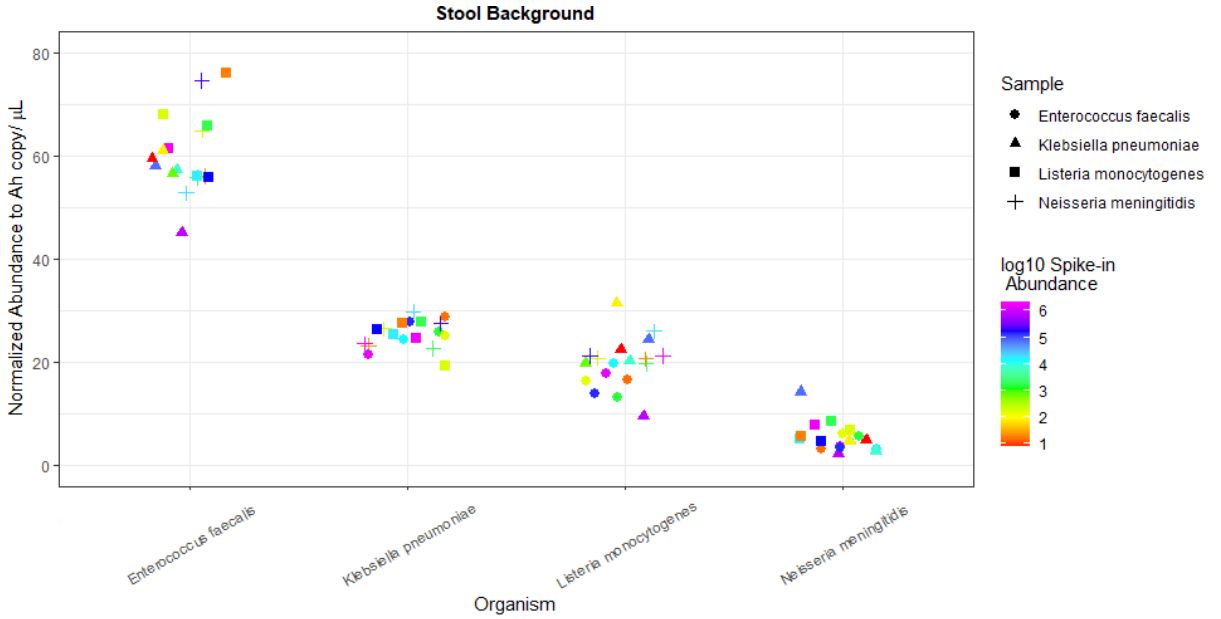
Stool use case

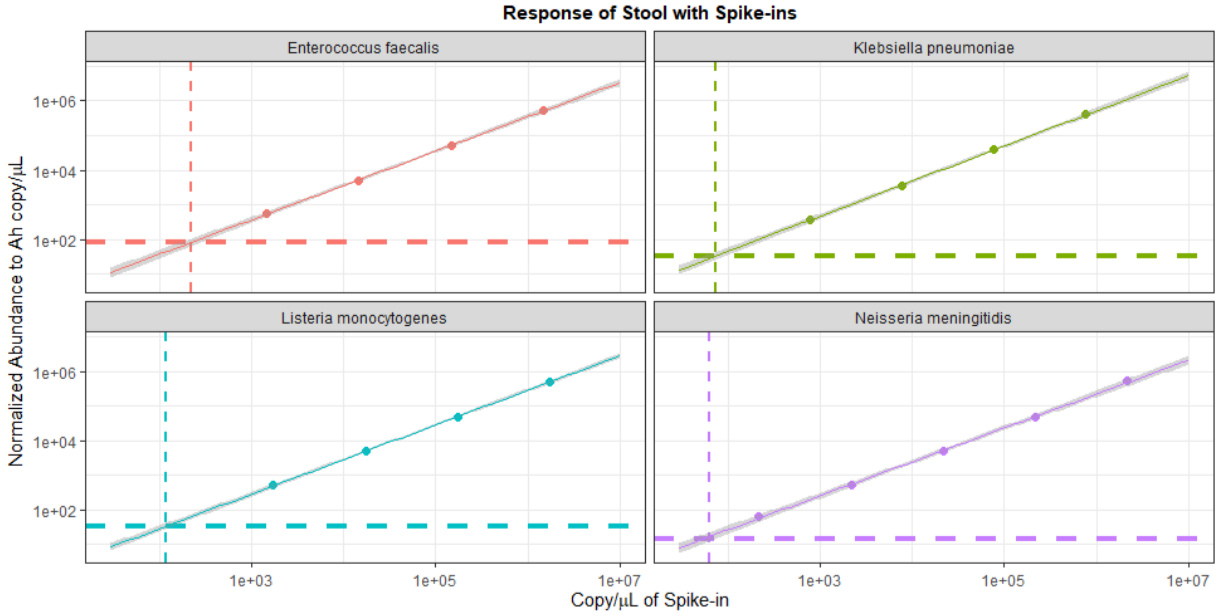
The stool spike-in samples were analyzed using a similar workflow to the CSF samples. Briefly, we added two internal control organisms and variably spiked with a log-dilution of select RM 8376 components to extracted stool DNA. Those samples were sequenced and analyzed to look for the LOD and response of the spike-ins. We plotted the background signal and above-background response for the samples (Fig. 3).

The background signal for the spike-in taxa was significantly higher in stool than for CSF, as will be shown later. We did not observe any significant changes in the background from high vs low abundance spike-in samples; however, we only included samples with spike-ins less than 10^4 copy/µL (n = 12 per taxon) to be consistent for comparisons to CSF in later discussions. Also, unlike the CSF data, the spike-in concentrations span the LOD. The linear regression and LOD calculations were tabulated and compared with those of CSF (Fig. 2, bottom).

The background data showed a clear difference in the LOD for each taxon (Fig. 3, top). We expected *Enterococcus* species could be present in the stool and generate significant background signals for that spike-in, and this was the case. The other three spike-ins also showed background signal. Given the broad diversity and high abundances of stool microbes, this result was not particularly surprising.

The Kraken2 workflow results are shown in Fig. S2. The response functions were similar between Kraken2 and Centrifuge. However, the LODs for the Kraken2 workflow for these taxa were lower than those for Centrifuge:default, indicating that alternative workflows/databases may improve the performance. While Kraken2 may outperform Centrifuge for this sample set, both may be suitable for use if they meet the LOD needed for clinical utility; other factors including performance of a specific panel of organisms, speed, and ease of use may factor into the use of one/both/neither. We will explore this further in attempting to discriminate the strains of *E. coli*.





Organism	Bkgd normAb to Ah mean (sd) copy/µL	Linear Regression m (CI)	Linear Regression b (CI) copy/µL	LOD mean (CI) copy/µL	
E. faecalis	61.2 (7.1)	0.99 (0.04)	0.38 (0.18)	221	(107)
K. pneumoniae	25.6 (2.7)	1.01 (0.04)	0.43 (0.22)	73	(37)
L. monocytogenes	20.5 (4.5)	1.00 (0.03)	0.28 (0.07)	119	(33)
N. meningitidis	5.6 (2.8)	1.00 (0.03)	0.22 (0.11)	62	(31)

FIG 3 Individual organisms were spiked into human stool at different concentrations ranging from approximately 10^6 to 10^1 copy/µL. (Top) The background signal for each organism was assessed similar to the criteria used for CSF. The background signal for each taxon in stool samples was significantly higher than in CSF. This was expected due to the rich and diverse population of microbes found in stool, with some bias due to highly abundant taxa (e.g., *Enterococcus*). Using

(Continued on next page)

Fig 3 (Continued)

a taxon-specific assessment, differences in the performance were easily identified, and this information could be used to properly establish LODs. (Middle) The response for each organism was plotted as the normalized abundance to Ah vs. the spike-in concentration as described in Fig. 2. Here, both the magnitude and the differences in LOD by taxon were larger than with CSF, likely due to confounding signals from related taxa within the stool. (Bottom) Analytical performance metrics for the Centrifuge classifier for stool. The same method was used as described in Fig. 2. As expected, the background signal and LODs for each taxon in stool samples were significantly higher than those for CSF. However, the linear regression for each organism had nearly identical parameters regardless of the background, suggesting that for any given taxon, the workflow response is independent of the sample composition.

Demonstrating the detection specificity of commensal vs pathogenic *E. coli*

Another test for mNGS is to demonstrate the ability to discriminate organisms such as *C. difficile* and various pathogenic strains of *E. coli,* etc. from commensal and/or nonpathogenic near-neighbor species or strains. Isolates can be sequenced deeply and assembled to paint a clear picture; however, this is time-consuming and expensive and may not result in improved clinical outcomes. Instead, mNGS-based systems that can relatively rapidly identify/discriminate pathogenic strains from whole stool directly, without isolation, would have clinical utility.

We tested the effect of using different databases with the same classifier and data to see if the workflows would perform differently. For this demonstration, the classifier Centrifuge was used with its default and the Web of Life (WoL) (20) databases, keeping all other parameters the same.

The goal of the analysis was to detect and differentiate *E. coli* sequences generated from the addition of commensal K12 strain DNA vs two pathogenic strains in RM 8376 (an O157:H7 and O104:H4 strain). Since the workflow can report on both species- and strain-level detection, we set up our analysis to report on *E. coli* at the species level and strains *E. coli* O157:H7 and *E. coli* O104:H4. All samples should contain *E. coli*, with only the pathogenic strains giving positive signals for their respective strain when present. We also chose to report the signal:background (similar to how others have normalized reporting [16]) because the background for *E. coli* in this sample was relatively high, making species- and strain-level normalized abundances appear disparate. (Fig. 4).

The results showed that the classifier Centrifuge:default correctly identified the addition of *E. coli* DNA above background in all samples and only identified the correct pathogenic strains when present. In comparison, using the same workflow but with the WoL database did not result in the discrimination of K12 *E. coli*. The K12 *E. coli* DNA gave a strong positive response for both pathogenic strains. Both pathogenic strains gave strong positive signals for their correct strain ID, with weak positive signals for the other. While the default database was able to discriminate pathogen vs nonpathogenic strains, the WoL database failed the basic discrimination test.

In addition to testing a single classifier with two databases, we also evaluated the performance of the Kraken2 workflow with the default database. This workflow could not identify the O104:H4 strain despite its inclusion in the database, with the sample containing nearly 10^6 genome copy/µL registering 0 reads. The workflow did show good discrimination of commensal and O157:H7 strains. However, this clear lack of response for O104:H4 (Fig. S3) necessitates a different approach to meet the needs of this (*E. coli* discrimination) analysis. These data highlight the impact that database selection can have on a workflow, and that taxonomic classifiers' utility is strongly tied to this critical feature. Furthermore, differences between Kraken2 and Centrifuge performance further highlight the need to examine the entire panel of target taxa when selecting a metagenomic workflow. While this does not represent an exhaustive list of potential pathogenic targets, it shows that the metagenomic workflow can discriminate these pathogenic and commensal strains. Hence, both the materials and the methodologies outlined here enabled a relatively simple and rapid assessment of fitness-for-purpose without employing any clinical specimens.

Normalized Abundance @ 10^5 copy/µL Spike-In

		Signal / Background		
		species	strain	
Workflow	Sample	*E. coli*	O157:H7	O104:H4
Centrifuge: Default	*E. coli* K12	6.3	0.6	0.4
	E. coli O157:H7	9.1	23.5	0.5
	E. coli O104:H4	7.8	0.8	60.3
Centrifuge:Web Of Life	*E. coli* K12	3.9	**5.4**	**2.5**
	E. coli O157:H7	6.4	57.9	**1.0**
	E. coli O104:H4	5.9	**1.2**	7.6
Kraken2: Default	*E. coli* K12	18.3	0.8	nd
	E. coli O157:H7	79.8	18.4	nd
	E. coli O104:H4	24.9	0.8	**nd**

Legend: TP FP TN FN

FIG 4 Analytical response of three *E. coli* strains spiked into stool at approximately 10^5 copy/µL using the classifier Centrifuge with two databases and Kraken2. For ease of comparison, the signal/background is the measured normalized copy number concentration to the background concentration for each taxon. The default database correctly identified all three samples as having *E. coli* and correctly discriminated the pathogenic strains from the commensal strain and each other. The WoL database also correctly identified all three samples as *E. coli*; however, the sample with the K12 strain was identified as having both O157:H7 and O104:H4 strains present. There was also a positive signal crossover between pathogenic strains, although the off-target signal was significantly lower than the on-target, which would be easily filtered. The Kraken2:default workflow could not detect the *E. coli* O104:H4 strain (noted as "nd") but could correctly discriminate the O157:H7 strain from the commensal and O104:H4 strains. The response curves for each sample and strain can be found in Fig. S3.

Stool analysis final thoughts

It is important to verify the workflow performance for each taxon because it is critical to identify potential challenges before moving forward with clinical samples. The workflow response should not be extrapolated to other taxa (or other taxonomic classes). Though similar organisms may share similar performance, and performance from one organism may inform experimental design for another, the myriad factors previously discussed may improve or degrade detection and/or discrimination relative to other taxa.

Effect of different sample types on the analytical performance of individual organisms

Examination of the linear models of each taxon revealed that the slopes and intercepts for individual taxa did not change between CSF and stool samples for the Centrifuge:default (or Kraken2:default, as shown in the SI) workflows (DNA library preparation through bioinformatics steps). This suggests that a given organism will produce the same analytical response when analyzed with a single mNGS workflow regardless of the sample matrix. Thus, the "agnostic diagnostic" theory for mNGS-based systems may be extended from the ability to identify any target within a sample to also include the analytical response of an organism across different sample types. However, we examined a limited sample set (i.e., CSF & stool), so additional work examining more sample types

and complete workflows (i.e., including sample handling, and DNA extraction) will be needed to test this hypothesis.

In contrast, we observed that the LODs for the spiked-in taxa were approximately 250- to 900-fold higher in stool than CSF with our workflow. This was expected because measurement sensitivity is limited by the background signal, which varies by sample type. In this study, CSF presents a relatively low microbial biomass background, while stool contains a significantly larger (and more diverse) microbial population. Furthermore, specific organism LODs will be impacted by the variety of compositions found across clinical samples. Application-specific LODs should always be established experimentally for each organism.

Conclusions

RM 8376 is a bacterial pathogen DNA-based Reference Material (RM) that supports metagenomic analytic performance assessments. We demonstrated that RM 8376 can be spiked into a background of DNA from other clinically relevant samples to evaluate the analytical performance of metagenomic workflows from sequencing library preparation and sequencing through bioinformatics processing. Thus, this material is a critical component within the suite of reference materials needed to assess the performance at different steps in the metagenomics workflow.

We demonstrated the utility of DNA spike-ins for rapid metagenomic analytical performance assessment in samples differing in complexity by generating quantitative LODs and response functions and generating data that would lead to workflow optimization. For both low-diversity CSF samples and high-diversity stool samples, taxon-specific LODs were generated. As expected, the LODs for spike-in organisms were higher in stool than in CSF. The analysis of CSF samples showed similar detection below 1 genome copy/µL for all taxa, with similar analytical performance of the sequencing and informatics steps to previous work.

Within the framework of generating data supporting clinical utility, this approach generates an actionable pass/fail result for continued testing or redevelopment. If a workflow does not perform as expected on a DNA-only test set, it will likely perform worse when tested with whole-cell controls because additional variables (i.e., sample storage and nucleic acid extraction) negatively impact the workflow. Additional whole-cell reference materials and characterization would be required for those upstream steps, and a correction factor could be applied to the results generated here. Hence, DNA-based reference materials, such as RM 8376, support rapid analytical performance testing of metagenomic workflows.

RM 8376 includes several near-neighbor organisms to further assess metagenomic workflows for specificity. Different strains of *E. coli* were used and highlighted two key factors in workflow assessment. First, classifier and database selection impact the analytical performance. Of the three workflows tested (Kraken:default, Centrifuge:default, and Centrifuge:WebOfLife), only one could achieve the desired strain-level discrimination. Ideally, one would confirm the target taxa are contained within the database; however, even when this is the case (Web of Life), the classifier itself may not yield the desired performance/result. And second, each of the taxa, even down to the strain level, will have specific performance characteristics. The analytical performance results in this work should not be extrapolated to other taxa, though the results may inform workflow developers and reviewers on the capabilities of a workflow and serve to identify extraordinary claims requiring additional investigation.

Metagenomic platforms hold great promise to fill a much-needed void for an "agnostic diagnostic." Typically, this reference has referred to the system being target-/pathogen-agnostic; however, metagenomic workflows may go even further from an analysis standpoint and also be sample-agnostic—the signal/response of an individual organism may not depend on the background sample composition. Individual organism response functions for the CSF and stool spike-in data showed remarkable similarity, despite organism LODs differing by several orders of magnitude. However, only two

sample types were explored, evaluating only the detection of DNA (i.e., not whole cells). We plan to study this further.

Looking ahead, tackling the challenge of transitioning metagenomics from a proof-of-concept laboratory test into regulated application requires careful step-by-step characterization and assessment of the entire metagenomic workflow. This requires a framework of validation studies, each utilizing reference materials that have been designed to assess the performance covering each step in the workflow. In the work shown here, we highlighted the utility of DNA-based reference materials to characterize the analytical performance of taxonomic classifiers using two use cases. Building on the evidence found here, further studies employing *in silico* reference data could provide additional support for rapid workflow development and performance assessment. Reference data (i.e., genomes or raw sequencing reads) would provide the ground truth for such experiments. Additionally, whole-cell reference materials would allow testing of additional experimental parameters including sample collection/storage and extraction. Together, this proposed suite of standards, combining reference materials and data, will help provide the experimental evidence for the translation of metagenomics into the regulated application space.

MATERIALS AND METHODS

Cerebrospinal fluid (CSF)

The frozen CSF (BioIVT) obtained from five healthy volunteers was thawed on ice and extracted using the QIACube (Qiagen) with the Qiagen Blood and Tissue Kit (Cat #69504) and pooled. The CSF was separated into 12 180-µL aliquots, with elution of 100 µL each, pooled together totaling 1.2 mL at 0.023 ng/µL determined using the Denovix HS DNA kit (Wilmington, DE USA).

A 30 µL aliquot of post-extraction DNA sample mixture was removed for sequencing to serve as the CSF negative control (no spike-ins or internal controls). *A. hydrophila* and *L. pneumophila* from RM 8376 (NIST, Gaithersburg MD, USA) were added as internal controls to the post-extraction DNA sample mixture at 1E4 genome copy/µL and 1.4E3 genome copy/µL, respectively. These concentrations were selected to ensure sufficient signal for normalization, but were not optimized. About 30 µL of this was removed for sequencing as the CSF with internal controls only.

Four components of RM 8376 were used as spike-ins to the CSF DNA: *K. pneumoniae* (starting concentration 7.68E6 copy/µL), *N. meningitidis* (starting concentration 21.7E6 copy/µL), *E. faecalis* (starting concentration 14.8E6 copy/µL), and *L. monocytogenes* (starting concentration 17.4E6 copy/µL). Sample sets for each spike-in were generated using 1:10 serial dilutions (i.e., 5 µL *K. pneumoniae* stock into 45 µL CSF DNA and then 5 µL of that mix into 45 µL of CSF DNA, etc.) for six dilution sets of each spike-in. With four sample sets, two CSF controls, and a DNA elution buffer control, a total of 27 samples were then prepared for DNA shotgun sequencing.

Stool

Two 1 mL vials (containing 100 mg each) of homogenized candidate fecal reference material (BioIVT, Westbury NY, USA) were split into four aliquots each (250 µL) for extraction of genomic DNA using the Zymo Quick-DNA Fecal/Soil Microbe kit (KIT #D6012, Zymo Research) with 40 minutes of bead-bashing at maximum speed (Vortex Genie, Scientific Industries, Inc., USA). The eight DNA extractions were pooled and quantified at 7.0 ng/µL. The total volume was brought to 2.5 mL using 1 × TE (Sigma-Aldrich, USA, Cat# 93302–500 mL). Of this, 30 µL was withheld for sequencing as a negative control.

Two internal controls were added to the pooled DNA. Approximately 25 µL of *A. hydrophila* DNA and 2.5 µL of *L. pneumophila* DNA were used to achieve 10^5 genome copy/µL and $1.4{\cdot}10^4$ genome copy/µL, respectively. Another 30 µL of this was withheld for sequencing as stool DNA with internal controls.

Four components of RM 8376 were used as spike-ins to the stool DNA: *K. pneumoniae* (starting concentration 7.68E6 copy/µL), *N. meningitidis* (starting concentration 21.7E6 copy/µL), *E. faecalis* (starting concentration 14.8E6 copy/µL), and *L. monocytogenes* (starting concentration 17.4E6 copy/µL). Sample sets for each spike-in were generated using 1:10 serial dilutions (i.e., 5 µL *K. pneumoniae* stock into 45 µL stool DNA and then 5 µL of that mix into 45 µL of stool DNA) for six dilution sets of each spike-in.

For the *E. coli* experiments, two components of RM 8376 were used as spike-ins to the stool DNA: *E. coli* O157:H7 (starting concentration 8.84E6 copy/µL) and *E. coli* O104:H4 (starting concentration 8.86E6 copy/µL). Sample sets for each spike-in were generated using 1:10 dilutions as previously described for six dilution sets of each spike-in. The *E. coli* K12 DNA was obtained from Sigma-Aldrich (MBD0013, approximately 10 ng/µL, estimated starting concentration 1.7e6 copy/µL). The five-sample set consisted of first diluting the stock 1:20 and then four additional 1:10 dilutions.

Sequencing protocols

Samples were prepared using the New England Biolabs (Ipswich MA, USA) NEBNext Ultra II FS kit (NEB #E7805, E6177) and the NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) (E7600S) following the manufacturer's instructions including 12 cycles of PCR. For CSF samples, the DNA concentration was measured using the Denovix dsDNA HS for each sample after library preparation and found to range between 34 ng/µL and 81 ng/µL for the spiked-in samples, 17.5 ng/µL for the CSF negative control, 80 ng/µL for the CSF with internal controls, and 0.9 ng/µL for the elution buffer. For stool samples, each sample was analyzed for DNA length and concentration using the TapeStation D5000 (Agilent, Santa Clara, CA) and pooled equally. For sequencing, the sample pool was diluted to 12 pmol/L.

The DNA was sequenced using an Illumina MiSeq (San Diego, CA) with a MiSeq Reagent Kit v2 (300 cycle) as 2 × 150 paired-end reads (Cat #MS-102–2002) following the manufacturer's protocol, including PhiX controls at 12.5 pmol/L.

Data analysis

Raw data

The CSF data typically contained approximately (500 k to 1.0M) paired-end reads after quality filtering and before host removal. The stool samples contained approximately (600 k to 2.0M) reads per sample after quality filtering and before host removal.

Bash scripts

There were three bash scripts used to analyze the fastq files (SI-Bash Scripts). In Script #1, samples were de-hosted using Bowtie2 against the human genome assembly GRCh38. In Script #2, we performed quality filtering (bbduk), taxonomic classification using Centrifuge with the default database (21), Centrifuge with the Web of Life database (20), and Kraken2 with the default database (22). Centrifuge outputs were converted to a Kraken-style output using Centrifuge-kreport. (SI-Bash Scripts)

The third bash script queried the kreports for reads of the specific taxa of interest using a list of organisms of interest and storing this as a < sample > .csv file.

The <sample > .csv files were then used as inputs in an R script to perform final normalized abundance calculations and plotting.

Normalizing abundance

Normalized abundance was selected over relative abundance, reads per million (RPM), and reads per kilobase per million reads (RPKM). Because the normalization factor (Ah concentration) was known, the response from each taxon of interest can be quantified *independently* of other taxa (10, 23). This would not be the case with the other metrics, where bias from every organism impacts the reported metric.

We calculated the normalized abundance for each taxon in each sample relative to *Aeromonas hydrophila* (Ah). The internal control concentration of Ah in CSF (10^4 copy/µL) and stool (10^5 copy/µL) was held constant for each sample type. The formula to calculate the normalized abundance was:

$$NormAb_i = \frac{Reads_i}{Reads_{Ah}} C_{Ah}, \quad (1)$$

where $NormAb_i$ is the normalized abundance of species i, $Reads_i$ are the assigned reads by the taxonomic classifier to species i, and C_{Ah} is the concentration of Ah.

In the experimental design, we also included *Legionella pneumophila* (Lp) as a second internal control. The Ah/Lp ratio (approximately 10) could be used to check for contaminating Ah and/or Lp that would impact normalization (>10 indicating background Ah and <10 indicating background Lp). In this study, Ah and Lp background levels were not significant compared to the internal control abundance.

Post-sequencing analysis

Minimum signal

For each taxon, the background normalized abundance was estimated using samples where (i) the taxon was not spiked-in and (ii) only samples with spike-in abundances below 10^4 copy/µL. For each taxon, this resulted in approximately 12 samples. This ensured that the sample compositions were similar and not overwhelmed by spike-in reads.

The background normalized abundance mean and standard deviation (sd) were calculated for each taxon. We set the minimum signal for finding the LOD (mean + 3 sd).

Linearity

For each taxon, we filtered data for spike-ins with normalized abundance above the minimum signal. We fit the log-transformed data to a linear model log (NormAb) = m log (spike-in) +b. The 95% confidence interval on the slope and intercept was estimated using Student *t*-test.

Limit of detection (LOD)

For each taxon, we established the LOD by solving the linear model at the minimum signal for the spike-in concentration. The LOD confidence interval was estimated using the expanded uncertainty with the coefficient of variabilities of the slope and intercept of the model.

ACKNOWLEDGMENTS

Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

AUTHOR AFFILIATION

[1]Complex Microbial Systems Group, Biosystems and Biomaterials Division, Materials Measurements Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

AUTHOR ORCIDs

Jason G. Kralj http://orcid.org/0000-0003-4565-1176

AUTHOR CONTRIBUTIONS

Jason G. Kralj, Conceptualization, Data curation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review and editing | Stephanie L. Servetas, Conceptualization, Project administration, Writing – review and editing | Samuel P. Forry, Conceptualization, Writing – review and editing | Monique E. Hunter, Investigation, Resources | Jennifer N. Dootz, Investigation, Resources | Scott A. Jackson, Supervision, Writing – review and editing

DATA AVAILABILITY

The raw, dehosted (human) sequencing data can be found at the NCBI SRA under PRJNA1074773.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (Spectrum02806-24-s0001.pdf). Fig. S1 to S3; supplemental data.

REFERENCES

1. Gauthier NPG, Chorlton SD, Krajden M, Manges AR. 2023. Agnostic sequencing for detection of viral pathogens. Clin Microbiol Rev 36:e0011922. https://doi.org/10.1128/cmr.00119-22
2. Govender KN, Street TL, Sanderson ND, Eyre DW. 2021. Metagenomic sequencing as a pathogen-agnostic clinical diagnostic tool for infectious diseases: a systematic review and meta-analysis of diagnostic test accuracy studies. J Clin Microbiol 59:e0291620. https://doi.org/10.1128/JCM.02916-20
3. Kralj J, Tourlousse D, Hunter M, Romsos E, Toman B, Vallone P, Jackson S. 2022. Reference material 8376 microbial pathogen DNA standards for detection and identification. US DOC, NIST Publications. Available from: https://www.nist.gov/publication
4. Kralj JG, Servetas SL, Forry SP, Jackson SA. 2020. Considerations for performance metrics of metagenomic next generation sequencing analyses. bioRxiv. https://doi.org/10.1101/2020.12.17.423212
5. Venkataraman A, Parlov M, Hu P, Schnell D, Wei X, Tiesman JP. 2018. Spike-in genomic DNA for validating performance of metagenomics workflows. Biotechniques 65:315–321. https://doi.org/10.2144/btn-2018-0089
6. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2017. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res 45:e23. https://doi.org/10.1093/nar/gkw984
7. Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, Ovaskainen O. 2020. SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol Ecol Resour 20:256–267. https://doi.org/10.1111/1755-0998.13057
8. Li M, Tyx RE, Rivera AJ, Zhao N, Satten GA. 2022. What can we learn about the bias of microbiome studies from analyzing data from mock communities? Genes (Basel) 13:1758. https://doi.org/10.3390/genes13101758
9. Kumar MS, Slud EV, Okrah K, Hicks SC, Hannenhalli S, Corrada Bravo H. 2018. Analysis and correction of compositional bias in sparse sequencing count data. BMC Genomics 19:799. https://doi.org/10.1186/s12864-018-5160-5
10. McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing experiments. Elife 8:e46923. https://doi.org/10.7554/eLife.46923
11. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019. Establishing microbial composition measurement standards with reference frames. Nat Commun 10:2719. https://doi.org/10.1038/s41467-019-10656-5
12. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, Kawli T, Christians FC, Venkatasubrahmanyam S, Wall GD, Cheung A, Rogers ZN, Meshulam-Simon G, Huijse L, Balakrishnan S, Quinn JV, Hollemon D, Hong DK, Vaughn ML, Kertesz M, Bercovici S, Wilber JC, Yang S. 2019. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol 4:663–674. https://doi.org/10.1038/s41564-018-0349-6
13. Bloom SM, Mafunda NA, Woolston BM, Hayward MR, Frempong JF, Abai AB, Xu J, Mitchell AJ, Westergaard X, Hussain FA, Xulu N, Dong M, Dong KL, Gumbi T, Ceasar FX, Rice JK, Choksi N, Ismail N, Ndung'u T, Ghebremichael MS, Relman DA, Balskus EP, Mitchell CM, Kwon DS. 2022. Cysteine dependence of *Lactobacillus iners* is a potential therapeutic target for vaginal microbiota modulation. Nat Microbiol 7:434–450. https://doi.org/10.1038/s41564-022-01070-7
14. Eisenstein M. 2020. The skin microbiome. Nature New Biol 588:S209. https://doi.org/10.1038/d41586-020-03523-7
15. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, et al. 2011. Enterotypes of the human gut microbiome. Nature New Biol 473:174–180. https://doi.org/10.1038/nature09944
16. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, Stryke D, Pham E, Fung B, Bolosky WJ, Ingebrigtsen D, Lorizio W, Paff SM, Leake JA, Pesano R, DeBiasi R, Dominguez S, Chiu CY. 2019. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res 29:831–842. https://doi.org/10.1101/gr.238170.118
17. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, Federman S, Stryke D, Briggs B, Langelier C, et al. 2019. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N Engl J Med 380:2327–2340. https://doi.org/10.1056/NEJMoa1803396
18. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, Pamp SJ. 2016. Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. mSystems 1:e00095-16. https://doi.org/10.1128/mSystems.00095-16

19. Gu W, Deng X, Lee M, Sucu YD, Arevalo S, Stryke D, Federman S, Gopez A, Reyes K, Zorn K, Sample H, Yu G, Ishpuniani G, Briggs B, Chow ED, Berger A, Wilson MR, Wang C, Hsu E, Miller S, DeRisi JL, Chiu CY. 2021. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. Nat Med 27:115–124. https://doi.org/10.1038/s41591-020-1105-z
20. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, et al. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun 10:5477. https://doi.org/10.1038/s41467-019-13443-4
21. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 26:1721–1729. https://doi.org/10.1101/gr.210641.116
22. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0
23. Zaramela LS, Tjuanta M, Moyne O, Neal M, Zengler K. 2022. synDNA-a synthetic DNA spike-in method for absolute quantification of shotgun metagenomic sequencing. mSystems 7:e0044722. https://doi.org/10.1128/msystems.00447-22