Development and Multicenter Assessment of a Reference Panel for Clinical Shotgun Metagenomics for Pathogen Detection

Donglai Liu
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
Teng Xu
Yunnan Agricultural University
Haiwei Zhou
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
Qiwen Yang
Peking Union Medical College Hospital Department of Clinical Laboratory
Xi Mo
Shanghai Children's Medical Center Affiliated to Shanghai Jiaotong University School of Medicine Institute for Pediatric Translational Medicine
Dawei Shi
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
Jingwen Ai
Huashan Hospital Fudan University Department of Infectious Diseases
Jingjia Zhang
Peking Union Medical College Hospital Department of Clinical Laboratory
Yue Tao
Shanghai Childrens Medical Center Affiliated to Shanghai Jiaotong University School of Medicine Institute for Pediatric Translational Medicine
Donghua Wen
Shanghai East Hospital Department of Laboratory Medicine
Yigang Tong
Beijing University of Chemical Technology
Lili Ren
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Pathogen Biology; Chinese Academy of Medical Sciences and Peking Union Medical College
Wen Zhang
Chinese Center for Disease Control and Prevention
Shumei Xie
Vision Medicals Center for Infectious Diseases

Weijun Chen
BGI PathoGenesis Pharmaceutical Technology
Wanli Xing
Tsinghua University School of Medicine;CapitalBio Technology Co., Ltd
Jinyin Zhao
Dalian GenTalker Clinical Laboratory
Yilan Wu
Guangzhou Sagene Biotech Co., Ltd
Xianfa Meng
Guangzhou Kingmed Diagnostics
Chuan Ouyang
Hangzhou MatriDx Biotechnology Co., Ltd
Zhi Jiang
Genskey Medical Technology Co., Ltd
Zhikun Liang
Guangzhou Darui Biotechnology
Haiqin Tan
HangzhoulngeniGenXunMinKang Biotechnology Co., Ltd
Yuan Fang
Dinfectome Inc
Nan Qin
Realbio Genomics Institute
Yuanlin Guan
Hugobiotech Co.,Ltd
Wei Gai
WillingMed Technology (Beijing) Co., Ltd
Sihong Xu
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
Wenjuan Wu
Shanghai East Hospital Department of Laboratory Medicine
Wenhong Zhang
Huashan Hospital Fudan University Department of Infectious Diseases
Chuntao Zhang
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
Youchun Wang（✉ wangyc@nifdc.org.cn）
National Institute of Food and Drug Control: China National Institute for Food and Drug Control
https://orcid.org/0000-0001-9769-5141

Research

Keywords: metagenomic assays, pathogen detection, shotgun pathogen metagenomics

Posted Date: February 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-208796/v1

Abstract

Background

Clinical shotgun metagenomics for pathogen detection has been used in diagnosing infectious diseases. However, its technical assessments have been limited to reference standards by individual labs, single experimental workflow and laboratory.

Results

Here we reported the design and development of a set of reference reagents and reporting metrics dedicated for clinical metagenomics by the National Institutes for Food and Drug Control (NIFDC) of China, and a joint evaluation study including 17 independent laboratories in cross-workflow and cross-platform settings. Our results showed that the performance of metagenomic assays was significantly impacted by the factors of microbial types, host context, and read depth, thus highlighting the importance to take these factors into consideration when designing reference reagents and benchmarking assays. Through this study, we found false positives to be a common challenge across centers, and considerable site and library effects to limited the assay's quantitative value. Our multicenter study also provided practical guidance of performance that laboratory-developed shotgun pathogen metagenomics tests should aim to detect microbes at 500 CFU/mL (or copies/mL) in clinically relevant host context (10^5 human cells/mL) within a 24h turn-around time, and with a read depth of 20M reads or lower. This collaboration work provided a unique resource comprising nearly 600 billion reads (>5Tb) for technical evaluation in clinical and regulatory settings.

Conclusion

We demonstrated that the performance of metagenomic assays was significantly impacted by the microbial type, the host context and read depth, which emphasizes the importance to consider these factors when designing reference reagents and benchmarking studies. Across sites, workflows and platforms, false positive reporting and considerable site/library effects were common challenges to the assay's accuracy and quantifiability. Our study also suggested practical guidance of performance that laboratory-developed shotgun pathogen metagenomics tests should aim to detect microbes at 500 CFU/mL (or copies/mL) in clinically relevant host context (10^5 human cells/mL) within a 24h turn-around time, and with a read depth of 20M reads or lower. This collaboration work provided a unique resource comprising nearly 600 billion reads (>5Tb) for technical evaluation in clinical and regulatory settings.

Introduction

Infectious diseases are a leading cause of death worldwide, attributable to a great variety of pathogens that belong to different microbial types. Rapid and precise identification of disease-causing pathogens is the key to effective clinical management but remains challenging in clinical settings [1, 2]. Conventional

diagnostics either rely on culturing, or require a presumptive diagnosis by the clinician prior testing. Recent advances in high-throughput sequencing and bioinformatics technologies have enabled rapid growth in the application of metagenomic testing to detect pathogens [3–7]. Importantly, the rapid identification of SARS-CoV-2, the causing agent of the COVID-19 pandemic, was highly attributable to the use of pathogen metagenomics assay [8–12].

Next-generation sequencing (NGS)-based assays have recently been widely applied in the fields of non-invasive prenatal testing and companion diagnostics for cancer treatment [13–15]. However, compared to these assays which analyze a limited number of genetic sites within the human genome, pathogen shotgun metagenomics faces unique challenges, as it involves a great variety of genomes from all organisms present in clinical samples [16–19]. The wide ranges of cellular and genomic characteristics of these organisms not only require that the assay can access all genetic contents (e.g. breaking all cellular structures), but also differentiate them (e.g. preventing false annotation of closely related species). So far, assessments of shotgun metagenomics for pathogen detection have been limited to reference standards by individual labs, single experimental workflow and laboratory. A multicenter evaluation study using a common set of dedicatedly designed reference reagents and performance metrics is hence highly desired, which is crucial for establishing performance standard, guiding proper result interpretation, further assay development and clinical adaptations, as well as providing valuable information in the regulatory perspective for such a newly emerging technology. Similar to the MicroArray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects, large-scale community efforts were coordinated for assessing the performance of microarray and RNA-seq technologies across laboratories, platforms and pipelines [20–23].

In this study, we described the development of a pathogen reference panel and reporting metrics dedicated for clinical shotgun metagenomics by the National Institutes for Food and Drug Control (NIFDC) of China which produces the majority of the country’s official reference reagents, and a joint evaluation study including 17 independent laboratories in cross-workflow and cross-platform settings. In total, over 580 billion reads and over 5 Tb of sequencing data were generated and studied. To our knowledge, the current study represents the largest effort to date to produce and analyze comprehensive reference datasets for pathogen metagenomics.

Results

Design of Pathogen Reference Panel, Reporting Metrics and Study Overview

To mimic the biological context of clinical specimens and enabling comprehensive assessments, we designed and constructed a panel of 9 pathogen reference (PR) reagents that covered 30 potentially pathogenic microorganisms of 5 different types (gram-/+ bacteria, fungi, and DNA/RNA viruses) and included 2x10^5/mL human cells as host background (table S1) [24, 25]. These 30 species comprised of 19 genera, with species intentionally chosen from the same genus to test the ability of the assays to

discriminate closely related microbes. It was also designed to include microorganisms with a wide range of genome sizes (from 0.7 Kb to 19.05 Mb) and GC contents (from 33.2% to 70.4%, Fig. 1A, table S2).

Among this reference panel of 9 PR samples, one served as control (pathogen reference control or PRC) and had no contrived microbes. The other 8 samples can be grouped into two sets (PRH1-PRH4 and PRL1-PRL4) or four pairs (*e.g.* PRH1 and PRL1). Each pair of PR samples comprised the same contrived microorganisms at two different titers. The one in the PRH group had microbes contrived at a 5-fold higher titer compared to their PRL counterpart (Fig. 1B). For instance, PR1H and PR1L both contained the same microorganisms (*Escherichia coli K1, Streptococcus pneumoniae, Cryptococcus neoformans, Echovirus 11, Herpes simplex virus 1, Human betaherpesvirus 5, Human herpesvirus 6B*), but each microorganism in PR1H was 5-fold greater in titer than in PRL1. Every reference reagent in the PR panel was verified by polymerase chain reaction (PCR)-based methods and distributed to 17 independent laboratory sites (centers C1-C17) for blinded metagenomic testing and bioinformatics analysis (Fig. 1C, Supplemental Methods). These laboratories employed various experimental procedures, bioinformatics pipelines, and sequencing platforms, which included 4 different sample preprocessing steps, two different nucleic acid extraction approaches, three different library preparation approaches, six different sequencing platforms, and four different bioinformatics methods.

To support quantitative assessments, the PR panel was tested undiluted and in 10 replicates at 1:10 dilution, except for the pathogen-free PRC sample. There were a total of 2,641 libraries sequenced (Fig. 1C, table S3, table S4), generating 587 billion reads and 5.51 TB of data. After blinded testing, each site was requested to report a list of microbes along with their mapped reads, as well as the corresponding raw sequencing data, for further independent meta-data analyses. Given the unique, agnostic nature of this assay, we assessed the results using performance metrics including the measures of *Recall*, *Precision*, and *F-score* to indicate the assay sensitivity, specificity, and overall accuracy (Fig. 1C).

Multicenter Evaluation of Clinical Metagenomics Using The PR Panel

We first evaluated the diagnostic performance across the 17 sites *F-scores* varied considerably across the 17 laboratory sites, with a range from 0.5-1.0 and an average of 0.81. Even though only 2 out of the 17 sites achieved an overall *F-score* of 1.0, 59% (10 sites) achieved an *F-score* of >0.83 (Fig. 2A, B). At sample level, nearly 40% of reached *F-scores* of over 0.9, nearly 70% were over 0.75, and only 4% were lower than 0.5 (fig. S1A). Visualization of the similarity in detected microbes demonstrated that results were clearly grouped by the reference sample, despite variances across sites (Fig. 2C).

Recall and *Precision* contributed differentially to the site-to-site variation in *F-score* (Fig. 2A, S1B). While *Recall levels* remained relatively consistent (average=0.88, range: 0.75-1.0), *Precision* varied significantly across sites (average=0.77, range: 0.45-1.0). Similar observations were made at the sample level (fig. S1A). To further dissect the cross-site variation in diagnostic performance, we analyzed the TP, FP, and FN results at each site, and found FP to be the most variable across sites, ranging from 0 to 35 counts at each site (fig. S1C), while TP and FN appeared to be relatively consistent, ranging from 21-30 and 0-9 counts, respectively. These results suggest that the overall assay performance across workflows

and sites was differentiated more by their ability to reduce false positive, than that to reduce false negative. Intriguingly, despite measuring two different aspects of assay performance, a significant positive correlation rather than trade-off was observed between *Recall* and *Precision* (*P*=0.013, fig. S1D).

Among different microbial types, RNA viruses appeared to be the most challenging type of microbes to detect, with an average *Recall* of only 0.71 across all sites, significantly lower than that of other pathogens. Both gram-positive and gram-negative bacteria had the highest *Recall* among all microbial types (0.96 and 0.94), followed by DNA viruses and fungi at 0.89 and 0.80, respectively (Fig. 2D). Similar *Recall* patterns were observed between PRH and PRL panels (fig. S2A). Most microorganisms at titers above 200 CFU/mL or copies/mL could be detected by >50% of the sites, despite some RNA viruses that were missed at even above 100,000 copies/mL (Fig. 2E). Among all the microorganisms in our panel, fungi and RNA viruses including *Echovirus 11, Human Respiratory Syncytical virus B, Human Parecho virus, Candida Albicans, and Candida Lusitaniae* were the most prevalent causes of the false-negative results (fig. S2B). In line with these findings, the ability to detect fungi and RNA viruses varied widely across sites whilst that for gram-positive and gram-negative bacteria was relatively consistent (Fig. 2E). These emphasized the importance of using a reference panel specifically designed for shotgun pathogen metagenomics to cover all microbial types, as many reference reagents for microbiome studies only include bacteria [26].

When assessing the technical turnaround time (TAT) of various workflows across all the sites. Fourteen sites had a TAT between 20-24 hours ranging from 15.4 to 40.0 hours. (Fig. 2F, table S5). The sequencing reaction took up the largest portion of the workflow, followed by library construction, nucleic acid extraction, and data analysis, with each constituting 66.6%, 14.3%, 5.9%, and 5.4% of the accumulated TATs, respectively (Fig. 2G, table S5).

Assay Sensitivity Depends on Microbe:Host Abundance Ratio

In the scenario of clinical specimens, pathogens almost always exist amid a variable abundance of host cells. Conventional molecular diagnostics, such as PCR-based assays, often work by detecting specific pathogens with limited interference from human or other microorganisms. Unlike these targeted assays, shotgun metagenomic assays involve unbiased analysis of all nucleic acid molecules within a sample. Thus, we proposed that not only the absolute pathogen abundance, but also the relative microbe:host abundance ratio may affect assay performance and should be built into the design of the reference reagents.

In our reference panel, as all samples in our panel included the same titer of human cells (2x10^5/mL), PRL therefore represented a 5-fold higher abundance than PRH in both absolute abundance and relative microbe:host abundance. On the other hand, 1:10 dilution of any sample represented a 10-fold decrease in absolute abundance, with the relative microbe:host abundance remained unchanged compared to its undiluted counterpart (Fig. 1B).

We compared the observed abundances (as indicated by the number of mapped reads) between PRHs and their PRL counterparts, and as well as between the undiluted and their diluted samples. We found that there was a 5-fold difference in median observed abundance between PRH and PRL, and 10-fold sample dilution did not result in lowered observed abundances (Fig. 3A). Consistent observations were made when bacteria, viruses, and fungi were analyzed separately (Fig. 3B). In agreement with these findings, a lowered relative abundance in PRL resulted in a lower *Recall* performance (Fig. 3C), while solely reducing the absolute abundance through sample dilution did not significantly affect the performance (fig. S3). These data showed that the relative microbe:host abundance ratio, but not absolute microbial abundance is a key determinant of assay sensitivity by pathogen shotgun metagenomics. Therefore, the limit of detection (LoD) of this assay should be assessed and defined with the relative abundance ratio, rather than the absolute microbial abundance as used for most conventional assays such as PCR-based diagnostics .

Intra- and Cross-site Comparison of Microbial Abundance

We set out to assess the potential of pathogen metagenomics in inferring the expected abundance from the number of reads. We defined the expected pathogen abundance in a sample as (pathogen genome size x pathogen titer) / (human genome size x human cell titer) x the total number of clean reads, and the observed abundance as the actual number of reads. We reasoned that for metagenomics to allow relative pathogen quantification, there should be a linear correlation between the observed and expected abundances. Linear regression analysis showed significant correlations between the expected and observed abundances, either when all the pathogens were analyzed as a whole or separately according to the types of microbes (P<0.001, Fig. 3D). A similar correlation was observed when the abundance of HPV contained in HeLa cells was used as an internal control for normalization (fig. S4). It was not unexpected that the observed abundance was generally lower than the theoretical expectation (fig. S5), which might reflect the loss of microbial nucleic acids during the experimental processes such as cell wall breaking. The significant correlation between the observed and expected abundances, along with the recovery of the microbe:host ratio, suggested the assay's ability for intra-site relative abundance measurement.

As microbial abundance was inferred by the fraction of mapped reads, we wondered if the numbers of mapped reads could be of indicative value across sites. Numbers of mapped reads per million (RPM) varied significantly across sites, with differences of up to two orders of magnitudes. Such a difference in RPM was not just a result of applying different techniques, as substantial variation was still observed when sites using similar technical workflows were grouped and compared (Fig. 3E). By analyzing each key technical component in the experimental procedures, our data revealed that host depletion and column-based extraction methods were associated with higher RPMs than other technical variables, whereas library preparation by ultrasound, endonuclease, or transposase did not show significant effects on RPM (Fig. 3F). Adaptation of a bead-beating step was associated with a lower RPM, in agreement with its negative correlation with *F-score* (Fig. 3G).

These results suggest that pathogen abundance can be inferred by RPM within each site, but without a way to normalize the “site effect”, cross-site comparisons provided limited information when conducting cross-center evaluation.

Library Effect Impacts Assay Variation

To understand the assay’s reproducibility, we took advantage of the large replicated datasets to measure the coefficient of variations (CV) of mapped reads at each site. The average CV was 0.65 and ranged between 0.12 and 1.10; and 75% of the sites had CVs below 0.5. This variation remained at a comparable level among sites that apply similar technical workflows (Fig. 4A). A host depletion step appeared to associate with a lower CV, which might be due to its higher RPM. While no differences were observed in other processes based on cell wall breaking and various nucleic acid extraction methods, we found that endonuclease- and transpose-based library preparation demonstrated the lowest and highest CVs of 0.4 and 1.0 (Fig. 4B), respectively, implying that this was an important step that introduced variances. When different types of microorganisms were analyzed individually, we found significantly higher CVs for fungal detection versus bacterial or viral detection (0.80, 0.51, and 0.54, respectively) (P<0.001, Fig. 4C).

To examine how much these fluctuations stemmed from read depth-dependent sampling noise, we performed random re-sampling from the pooled reads to represent such a variance and calculated the CVs of these simulated and experimental datasets. As shown in Fig. 4D, the overall CVs were significantly greater than the simulated CVs regardless of pathogen types. This difference in CV was consistent when each laboratory site or microbial type was assessed individually (Fig. 4C, E), suggesting that besides read depth-dependent sampling, other experimental variables also contribute considerably to the observed fluctuations in metagenomic results.

We then attempted to determine how much each of the read depth-dependent variance and other experimental variables contributed to the total variance. We identified a significant linear correlation with an adjusted R^2 of 0.48 and a slope of 0.8 (P<0.001, Fig. 4F), indicating that both read depth-dependent and experimental variances contributed significantly to the overall fluctuation. Among the potential experimental variances, a linear mixed model identified fungal pathogens and transposase-based library construction as significant contributors (table S6), which was consistent with our previous interpretations.

Our data suggest to take these variations into consideration when designing studies to evaluate such an assay. They also imply that precise quantitative measurement of pathogen abundance by shotgun metagenomics remains challenging until these variations are better understood and more sophisticated quantitative modeling is established.

Read-depth Dependency of Assay Recall

Next, we sought to understand how workflow-dependent technical variables may lead to varied site performance. Among the experimental steps, we found sample pre-treatment had a greater impact on assay performance compared to nucleic acid extraction, library preparation, or use of internal control. Preprocessing the samples with host cell depletion was significantly associated with improved F-scores (P<0.001). Unexpectedly, a bead-beating step designed for breaking cell walls did not always result in greater performance but was associated with an overall reduced F-scores (P<0.001). Different technical methods for nucleic acid purification and library preparation, different sequencing platforms, and the use of spike-in internal controls were not correlated with overall assay performance (Fig.3G, S6).
Using the Q30 score as a quality indicator, we also found that *F-score* and *Precision* (but not *Recall*) were positively correlated with higher sequencing data quality (P<0.05, Fig. 5A).

We further explored the impact of read depth on the assay performance. Although initially, the diagnostic performance improved as the read depth increased, further increase in data size beyond 10 million did not consistently result in higher scores (Fig. 5B). This observation supports the interpretation that the contribution of read depth to assay performance plateaus after a certain read depth. Leveraging our data which constitute the deepest sequencing of any sample set yet reported, we next set out to determine the optimal read depth by assessing how well the pathogens in our panel could be detected as a function of read depth. To allow raw data analyses, a CLARK-based pipeline was chosen for subsequent site-independent bioinformatics analyses as it demonstrated good performances in both simulated and experimental sequencing datasets [27-29] (fig. S7, and more details in Methods).

As shown in Fig. 5C, some pathogens could be detected with only 0.5 million total reads. For instance, site C12 achieved a full *Recall* of 1.0 at a read depth of 0.5 million in 6 of the 8 PR samples. Nonetheless, when considering the data from all sites, a read depth of 20 million reads enabled detection of most of the microorganisms in our panel, and above that point benefits by deeper sequencing decreased significantly (Fig. 5C).

We performed sub-analyses by different microbial types of bacteria, fungi, and viruses. The performance of fungal detection plateaued at 5 million reads, while the performance of bacterial and viral detection plateaued at 10 and 20 million reads, respectively (Fig. 5D). These observations were also in line with the interpretation that the sensitivity of pathogen metagenomics decreases as the size of the microbial genome decreases (virus<bacterium<fungus), as smaller genomes result in fewer numbers of nucleic acid fragments that can be sequenced.

These findings suggest read depth as a critical variable that impacts assay *Recall* when both developing and assessing metagenomic tests. Our results also indicate that although a metagenomic assay requires as few as 0.5 million reads per sample for pathogen detection under optimal conditions, in general, a read depth of 20 million was appropriate under most assay settings.

Assay Precision Is Challenged by Background Microbes

To better understand the causes of FP which we found earlier that substantially impacted assay performance, we further categorized the causes of FP results into four groups: cross-contamination, background microorganisms, species misclassification, and viral typing error (Fig. 6A). Among these, background microbes and misclassification of species were the leading causes of FP results (49% and 39%, respectively). These two causes also varied the most among sites (fig. S8).

We then sought to evaluate how much taxonomical misclassification could be attributed to the alignment algorithms. To ensure comprehensive microbial coverage, we included 100,000 reads each derived from a total of 108 species comprising 62 bacteria, 42 viruses, and 4 fungi into our simulated dataset and compared the alignment methods employed by the sites in this study (bwa, bowtie, and SNAP) [30-32] by measuring the percentages of simulated reads that were correctly or incorrectly classified. We found no significant differences at both the genus and species levels, or by microbial type (fig. S9), implying that the alignment method was not a critical performance-differentiating factor.

To gain more insights into FP results caused by background microorganisms, we compared the background patterns from all sites (Fig. 6B). Nine prevalent microorganisms were presented in >5 sites, while others were more site-specific (Fig. 6C). Background microbial patterns clustered partially depending on the methods of library construction and nucleic acid extraction used (Fig. 6B). These findings implied that microbial backgrounds can be derived from both common and workflow-specific sources and that addressing such issues to improve assay *Precision* would require site-dependent approaches.

Besides read counts, we also examined whether genome coverage and regional sequencing depth could be informative in discriminating TP from FN. We defined genome coverage as the fraction of genome covered by metagenomic sequencing, and the regional sequencing depth as the total sequencing length divided by the covered genomic fraction. TP results were associated with significantly lower regional sequencing depth and higher genome coverage (fig. S10), which was consistent with the fact that these microorganisms exist in the samples as full and uniform genomes. Similar observations were also made for FN results that were missed originally but discovered by our site-independent bioinformatics pipeline. All FP detections showed significantly lower levels of genome coverage. A significant increase in the level of regional depth was also found in background microbes, implying that they presented as genomic fragments instead of full microbial bodies or genomes in the samples.

Taking all these factors into consideration, we built a *Precision* filter that identified potential FP results through machine learning and applied it to the data derived from site C14, the site that had the highest level of FP results. Integrating RPM, RPM ratio (sample:control), and genome coverage, our method reduced FP results from 34 to 11 counts (Fig. 6D). Importantly, applying such a filter did not compromise *Recall*, suggesting a potential strategy for improving the precision of pathogen metagenomics.

Discussion

In this coordinated study with large-scale community efforts, nine reference samples that mimicked the context of clinical specimens of infectious diseases were profiled at 17 independent sites with various workflows. The data presented here provide one of the deepest assessments of pathogen metagenomics assay to date.

In current clinical settings, pathogen metagenomicsis mostly employed for acute and severe infections that cannot be diagnosed by conventional approaches such as culture or PCR-based assays [5, 7, 33-38]. TAT is highly critical in these scenarios [39]. In this multicenter study, we found that most of the pathogen metagenomic assays could be completed within 24 hours, with the sequencing step as the major time-consuming step. Given that longer read lengths generally require more reaction time in most sequencing platforms, these indicate that modifying the assay with shorter reads could be an effective and relatively straightforward strategy for achieving faster assay turn-around.

Lowering the cost by applying an appropriate read depth may allow the wider application of pathogen metagenomics [40]. Our analyses found that across most of the sites, 20 million reads were sufficient for pathogen identification. We also observed that under optimal assay conditions, pathogen detection could be achieved with as low as 0.5 million reads, indicating the potential of further cost reduction as the technology advances [41].

By including various types of pathogens as well as human cells in the design, our reference panel represent the common context of clinical specimens, such as cerebrospinal fluid, sputum, and bronchoalveolar lavage fluid, where infiltration of immune cells is often found under infection. Although sharing common characteristics, our reference samples may not precisely represent plasma specimens where human cell-free nucleic acids are believed to be more prevalent [3]. For instance, it remains to be determined whether assay variation would be influenced by the cell wall-breaking step, and how each library preparation method fits in the context of cell-free nucleic acids.

We demonstrated the abundance of human cells as a critical factor by showing that pathogen detection by metagenomics was directly affected by the relative abundance of pathogens to host cells. Our data supported a mathematical model in which a sample comprising 10^5/mL each of human cells and bacterial cells would only yield 0.1% of total reads mapped to the bacteria, assuming a human genome of 3Gb and a bacterial genome of 3Mb [16, 25]. When interpreting results from pathogen metagenomics, it is important to note that its sensitivity could be affected by the host nucleic acids, and therefore varied across samples [16]. Host nucleic acid depletion would therefore be highly valuable in improving the read yield by increasing the relative microbial abundance and eventually in lowering read depth and assay costs. A variety of approaches have been reported for host cell depletion [42-45]. However, cautious validation should be performed before applying these methods to ensure that pathogens are not biasedly “co-depleted” with the host cells during the process. Indeed, in a previous study, differential lysis could significantly reduce human cells but at the same time compromise detection of viral and certain bacterial pathogens [43].

Different from the other targeted assays, pathogen metagenomics is also unique in its potential for unbiased detection of novel pathogenic microbes, as was shown in the discovery of COVID-19. Such ability heavily depends on bioinformatics analysis to discriminate between novel and previously identified pathogens, as well as closely related ones, for instance, between SARS-CoV-2 and SARS-CoV. Evaluating such an unusual aspect of assay performance would require new designs of the reference reagents that represent potential novel species.

Data in this study included sequencing results generated from different platforms and workflows using the same set of reference samples. This information is a unique resource that could be valuable for the development and optimization of bioinformatics pipelines for rapid pathogen detection. Current bioinformatics pipelines mostly rely on the number of mapped reads for pathogen identification [46-48]. With our dataset, more sophisticated identification algorithms could be explored by integrating more variables, such as genome coverage and phylogenetic relationships, to improve specificity. These data also provide a general overview of the current performance of pathogen metagenomics, which could aid in establishing regulatory or technical references.

Conclusion

We have reported the design and development of a set of reference reagents and reporting metrics dedicated for shotgun genomics for pathogen detection which can help standardise the field of clinical metagenomics. By testing these reference reagents in a multicenter study that included 17 independent laboratories, we demonstrated that the performance of metagenomic assays was significantly impacted by the microbial type, the host context and read depth, and emphasized the importance to consider these factors when designing reference reagents and benchmarking studies. Moreover, across sites, workflows and platforms, we found false positive reporting and considerable site/library effects to be common challenges to the assay's accuracy and quantifiability. Our study also suggested practical guidance of performance for laboratory developed shotgun pathogen metagenomics assays.

Materials And Methods

Preparation and validation of the reference panel

Bacterial and fungal organisms were validated by Matrix-assisted Laser Desorption/ Ionization-Time Of Flight (MALDI-TOF, Bruker, Billerica, MA), Vitek 2 (bioMérieux, Craponne, France), and a BioFire FilmArray Multiplex PCR System (bioMérieux, Craponne, France), and quantifitated by standard plate counts. Viral organisms were validated by Sanger sequencing and quantitated by droplet digital PCR (ddPCR). These microbes were then placcontrived into PBS solutions with $2x10^5$/ml of HeLa cells (ATCC) at indicated concentrations (table S1). In the PRH group, these microorganisms were studicontrived at 200-350,000 CFU/ml for bacterial and 400-10,000 CFU/ml for fungal pathogens, at 660-2,000,000 copies/ml for DNA viruses and at 140-3,500,000 copies/ml for RNA viruses to represent common ranges of clinical infection.

Comparison of bioinformatics pipelines

We used Mason (Mason – A Read Simulator for Next Generation Sequencing Data, v0.1.2) to generate simulated sequencing data for 108 microbial genomes, which including 62 bacterial, 42 viral, and 4 fungal microorganisms. A total of 100,000 single-end, 75bp reads were generated for each microbe and subjected to taxonomic identification by Centrifuge [27], Kraken [29], and CLARK [28] pipelines separately. Assessment of pipeline performance was performed at both the genus and species levels and also by microbial class. Sensitivity was inferred by the number of reads mapped specifically to the correct taxa; and specificity was inferred by the percentage of reads mapped specifically to the incorrect taxa. Statistical comparisons were as done by Wilconxon rank tests.

As Similar strategy was used for the comparing the son among alignment algorithms of BWA [30], Bowtie2 [31], and SNAP [32].

Correlation analysis of observed and theoretical abundances

Raw sequencing data from 17 sites were analyzed using the site-independent CLARK-based pipeline to obtain the observed abundance of each microorganism in each sample, except for RNA viruses. The theoretical abundance of a microorganism in a sample was proportional to the ratio of DNA of that microorganism and the size of sequencing data, calculated as below:

$$\text{Theory abundance for microbie i} = \frac{\text{Copy}_i * \text{Genomesize}_i}{\text{Copy}_{human} * \text{Genomesize}_{human} + \sum_k \text{Copy}_k * \text{Genomesize}_k} * \text{data size}$$

Where Copy_i and Genomesize_i was the copy number and genome size of microorganism i in this sample, respectively. Human cell number Copy_{human} for each sample was constant to $2x10^5$, and human genome size $\text{Genomesize}_{human}$ was set at 3G. Subsequently, a linear regression model was used to estimate the correlations between the observed and theoretical abundances.

Analysis of Simulated and Observed Coefficient of Variants (CVs)

Fastq data from the 10 repeated replicates of each sample were merged, re-split randomly according to the original read depth (number of reads), and analyzed by the CLARK-based pipeline. The CVs were calculated based on the read numbers mapped to each microbes. This above process was repeated 10 times in order to obtain a total of 10 simulated CVs for each microbe. We used the average of 10 simulated CVs, as the expected CV resulted from read depth variation. A linear regression model was used to evaluate the contribution of the data size CVs to the observed CVs.

In addition, we used a linear mixed model to further evaluate whether the sequencing platform, library method, and class of microorganism affected the observed CV. The formula of the linear mixed model was defined as:

Cv_observed ~ Cv_datasize + Library Prep + Microbial Class + Platform + (1+Cv_observed|Center) where Center was a random effect, and the read depth CV (Cv_datasize), library preparation method (Library Prep), microbial class, and sequencing platform (Platform) were fixed effects.

Analysis of Read-depth Requirement for Pathogen Detection

Fastq data from the 10 repeated replicates of each sample were merged, and resampled to the desired read depth of 0.5M, 1M, 5M, 10M, 20M, 30M, and 50M total reads for each sample. The re-sampled data were analyzed by the CLARK-based pipeline and identification of a microorganism was defined by over 4 species-specific mapped reads in a sample. The recall performance was assessed at each indicated read depth for each site by sample or by microbial class.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Acknowledgments: Not applicable

Funding: This work was supported by National Science and Technology Major Project of China (2018ZX10102001).

Author contributions: Y.W and C.Z conceived, designed and supervised the experiments; D.L, T.X, H.Z, Q.Y, X.M and Y.W wrote the manuscript; D.S, J.A, J.Z, Y.T, D.W, Y.T, L.R, W.Z, S.X, W.C, W.X, J.Z, Y.W, X.M, C,O, Z.J, Z.L, H.T, Y.F, N.Q, Y.G, and W.G performed the experiments. All of the authors have read and approved the final manuscript. S.X, W.W and W.Z helped with design of experiments, supervising a specific platform and proof of the manuscript.

Competing interests: The authors declare no competing interests.

Availability of data and material: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Following complete publication, the data generated in this study will be made available to researchers through GISAID accession ID CNP0001292.

References

1. Barlam TF, Cosgrove SE, Abbo LM, MacDougall C, Schuetz AN, Septimus EJ, et al. **Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America.** *Clin Infect Dis* 2016, **62:**e51-77.
2. Liesenfeld O, Lehman L, Hunfeld KP, Kost G. **Molecular diagnosis of sepsis: New aspects and recent developments.** *Eur J Microbiol Immunol (Bp)* 2014, **4:**1-25.
3. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, et al. **Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease.** *Nat Microbiol* 2019, **4:**663-674.
4. Ye SH, Siddle KJ, Park DJ, Sabeti PC. **Benchmarking Metagenomics Tools for Taxonomic Classification.** *Cell* 2019, **178:**779-794.
5. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. **Actionable diagnosis of neuroleptospirosis by next-generation sequencing.** *N Engl J Med* 2014, **370:**2408-2417.
6. Simner PJ, Miller HB, Breitwieser FP, Pinilla Monsalve G, Pardo CA, Salzberg SL, et al. **Development and Optimization of Metagenomic Next-Generation Sequencing Methods for Cerebrospinal Fluid Diagnostics.** *J Clin Microbiol* 2018, **56**.
7. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. **Pathogen Genomics in Public Health.** *N Engl J Med* 2019, **381:**2569-2580.
8. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. **A novel coronavirus from patients with pneumonia in China, 2019.** *New England Journal of Medicine* 2020.
9. Bulut C, Kato Y. **Epidemiology of COVID-19.** *Turk J Med Sci* 2020, **50:**563-570.
10. WHO. **Latest rolling update: WHO characterizes COVID-19 as a pandemic.** [https://www.who.int]
11. Zhang H, Ai J-W, Yang W, Zhou X, He F, Xie S, et al. **Metatranscriptomic Characterization of COVID-19 Identified A Host Transcriptional Classifier Associated With Immune Signaling.** *Clin Infect Dis* 2020.
12. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. **SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.** *Cell* 2020, **181:**271-280.e278.
13. Weber S, Spiegl B, Perakis SO, Ulz CM, Abuja PM, Kashofer K, et al. **Technical Evaluation of Commercial Mutation Analysis Platforms and Reference Materials for Liquid Biopsy Profiling.** *Cancers* 2020, **12**.
14. Salk JJ, Schmitt MW, Loeb LA. **Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations.** *Nat Rev Genet* 2018, **19:**269-285.

15. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. **Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology.** *J Mol Diagn* 2015, **17:**251-264.
16. Westermann AJ, Gorski SA, Vogel. **Dual RNA-seq of pathogen and host.** *Nat. Rev. Microbiol.* 2012, **10:**618-630.
17. Consortium TIHiRN. **The Integrative Human Microbiome Project.** *Nature* 2019, **569:**641-648.
18. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. **The vaginal microbiome and preterm birth.** *Nat Med* 2019, **25:**1012-1021.
19. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. **Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.** *Nature* 2019, **569:**655-662.
20. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol* 2006, **24:**1151-1161.
21. Consortium SM-I. **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.** *Nat Biotechnol* 2014, **32:**903-914.
22. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nat Biotechnol* 2010, **28:**827-838.
23. Yu L. **RNA-Seq Reproducibility Assessment of the Sequencing Quality Control Project.** *Cancer Inform* 2020, **19:**1176935120922498.
24. Ahlbrecht J, Hillebrand LK, Schwenkenbecher P, Ganzenmueller T, Heim A, Wurster U, et al. **Cerebrospinal fluid features in adults with enteroviral nervous system infection.** *Int J Infect Dis* 2018, **68:**94-101.
25. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. **Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis.** *N Engl J Med* 2019, **380:**2327-2340.
26. Amos GCA, Logan A, Anwar S, Fritzsche M, Mate R, Bleazard T, et al. **Developing standards for the microbiome field.** *Microbiome* 2020, **8:**98.
27. Kim D, Song L, Breitwieser FP, Salzberg SL. **Centrifuge: rapid and sensitive classification of metagenomic sequences.** *Genome Res* 2016, **26:**1721-1729.
28. 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.
29. Wood DE, Salzberg SL. **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol* 2014, **15:**R46.
30. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.

31. Langmead B, Salzberg SL. **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
32. Matei Zaharia WJB, Kristal Curtis, Armando Fox, David Patterson, Scott Shenker, Ion Stoica, Richard M. Karp, Taylor Sittler‡∗. **Faster and More Accurate Sequence Alignment with SNAP.** *arxivorg/pdf/11115572* 2011**:**1-10.
33. Li M, Yang F, Lu Y, Huang W. **Identification of Enterococcus faecalis in a patient with urinary-tract infection based on metagenomic next-generation sequencing: a case report.** *BMC Infect Dis* 2020, **20:**467.
34. Hoffmann B, Tappe D, Höper D, Herden C, Boldt A, Mawrin C, et al. **A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis.** *N Engl J Med* 2015, **373:**154-162.
35. Gu L, Liu W, Ru M, Lin J, Yu G, Ye J, et al. **The application of metagenomic next-generation sequencing in diagnosing Chlamydia psittaci pneumonia: a report of five cases.** *BMC Pulm Med* 2020, **20:**65.
36. Fang M, Weng X, Chen L, Chen Y, Chi Y, Chen W, et al. **Fulminant central nervous system varicella-zoster virus infection unexpectedly diagnosed by metagenomic next-generation sequencing in an HIV-infected patient: a case report.** *BMC Infect Dis* 2020, **20:**159.
37. Huang Z, Zhang C, Li W, Fang X, Wang Q, Xing L, et al. **Metagenomic next-generation sequencing contribution in identifying prosthetic joint infection due to Parvimonas micra: a case report.** *J Bone Jt Infect* 2019, **4:**50-55.
38. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. **RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak.** *Emerg Microbes Infect* 2020, **9:**313-319.
39. Miao Q, Ma Y, Wang Q, Pan J, Zhang Y, Jin W, et al. **Microbiological Diagnostic Performance of Metagenomic Next-generation Sequencing When Applied to Clinical Practice.** *Clin Infect Dis* 2018, **67:**S231-s240.
40. Chiu CY, Miller SA. **Clinical metagenomics.** *Nat Rev Genet* 2019, **20:**341-355.
41. Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, et al. **Efficient depletion of host DNA contamination in malaria clinical sequencing.** *J Clin Microbiol* 2013, **51:**745-751.
42. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, 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.
43. Nelson MT, Pope CE, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, et al. **Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles.** *Cell Rep* 2019, **26:**2227-2240.e2225.
44. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. **Improving saliva shotgun metagenomics by chemical host DNA depletion.** *Microbiome* 2018, **6:**42.
45. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. **Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples.** *Genome Biol* 2014, **15:**519.

46. Breitwieser FP, Baker DN, Salzberg SL. **KrakenUniq: confident and fast metagenomics classification using unique k-mer counts.** *Genome Biol* 2018, **19:**198.
47. Kang DD, Froula J, Egan R, Wang Z. **MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.** *PeerJ* 2015, **3:**e1165.
48. Corvelo A, Clarke WE, Robine N, Zody MC. **taxMaps: comprehensive and highly accurate taxonomic classification of short-read data in reasonable time.** *Genome Res* 2018, **28:**751-758.

Additional Information

H2: Supplementary Materials

Table S1. Design and composition of the pathogen panel.

Table S2. Pathogen panel genome characteristics.

Table S3. Sites and their corresponding technical parameters.

Table S4. Library numbers and data size in the study.

Table S5. Technical turnaround time of pathogen metagenomics workflow.

Table S6. A linear mixed model for identification of significant contributors of experimental variance.

Fig. S1. (A) Summary of performance metrics as measured by recall, precision, and F-score; (B) Site-by-site summary; (C) Site analysis of the occurrences of true positive (TP), false positive (FP) and false negative (FN); (D) Correlation analysis. Related to Fig. 2.

Fig. S2. (A) Heatmap showing the detected abundances at each site. Data were groups by microbial type, PRH and PRL. (B) Prevalence of FN results by microorganism. Related to Fig. 2.

Fig. S3. Site performance metrics between diluted and undiluted samples.

Fig. S4. Correlation between observed and expected abundances by microbial type after normalization with HPV virus as internal control.

Fig. S5. Comparison between the overall distribution of expected and observed microbial abundances.

Fig. S6. Correlations between recall (A) or precision (B) and key technical variables.

Fig. S7. Evaluation of three candidate pipelines for site-independent analysis.

Fig. S8. Count summary of all sites for each of the top four causes of false positive results.

Fig. S9. Evaluation of three candidate pipelines for site-independent analysis.

Fig. S10. Sequencing depth (left) and genome coverage (right) of the detected microbes in different result categories.

Figures

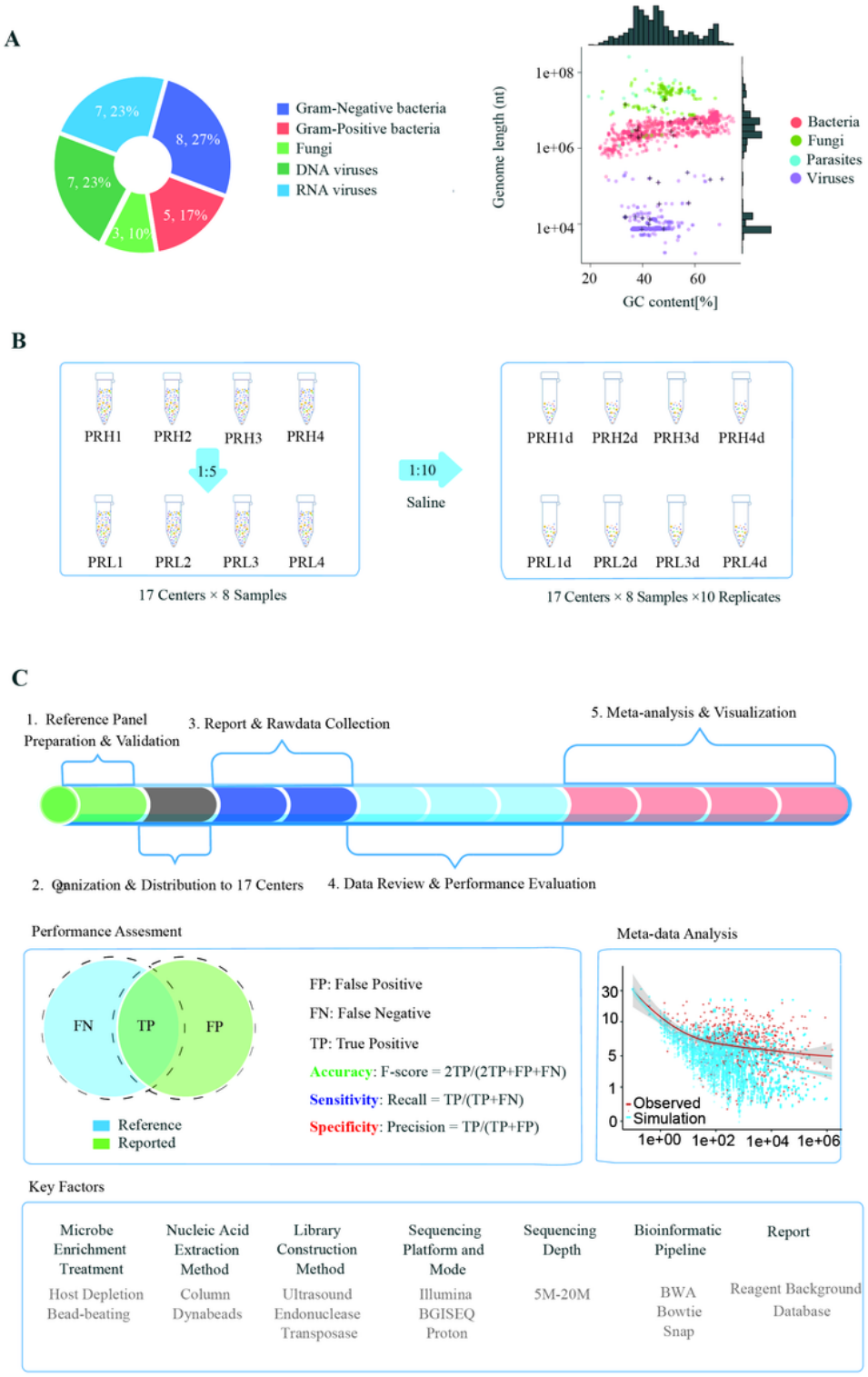

Figure 1

Design overview of the reference panel and multi-center study. (A) Thirty microorganisms from 19 genera were chosen to represent the diversity of microbial types (left panel), genome sizes, and GC contents (right panel). (B) Pathogen reference reagents (PRHs and PRLs) were prepared by contriving microbes at high (PRH) and low (PRL) titers with HeLa cells, respectively. PRH and PRL samples were diluted 1:10 and tested in ten replicates. (C) Overview of study design. Numbers (1-5) order the steps of analysis. F-score, Recall and Precision metrics were used to assess the performance of accuracy, sensitivity, and specificity. Key technical factors of the workflow were explored for their impact of assay performance. See also Table S1-S3, S5.

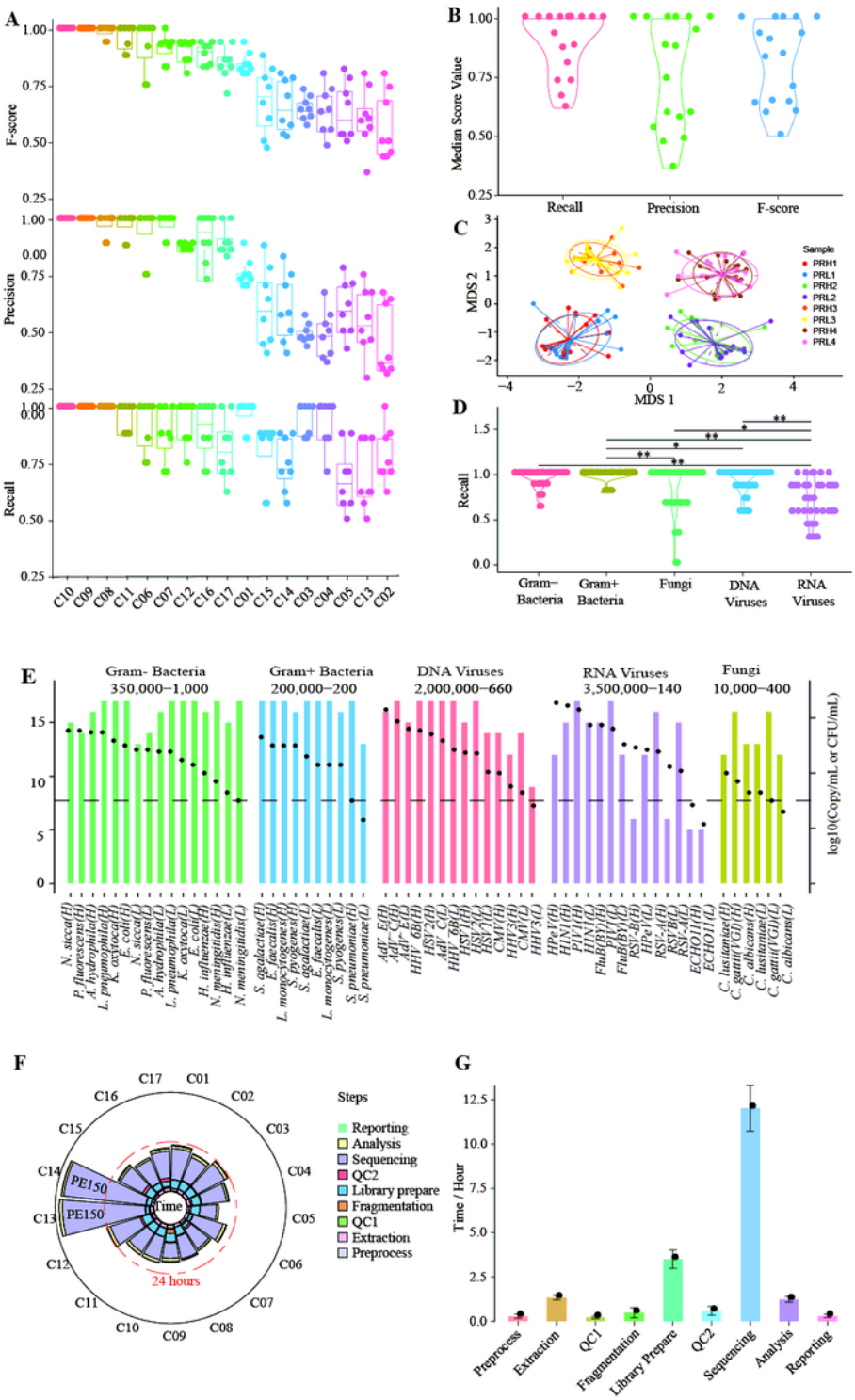

Figure 2

Assessment of assay performance across sites. (A) Summary of assay performance by site. (B) Summary of assay performance by measure. (C) Visualization of the similarity in microbial compositions across sites and reference reagents. A nMDS plot of a Bray-Curtis dissimilarity matrix was constructed from the species composition of each reference reagent at each site. (D) Summary of assay performance by microbial type, *, P<0.05, **, P<0.01. (E) Detection of each microorganism in the reference panel.

Bacteria and fungi were measured in CFU/ml, and viruses were measured in copies/ml. Ranges of abundances of each microbial type are displayed at the top. (F) Assay turn-around times by site. (G) Assay turn-around times by step. See also Fig. S1-S2.

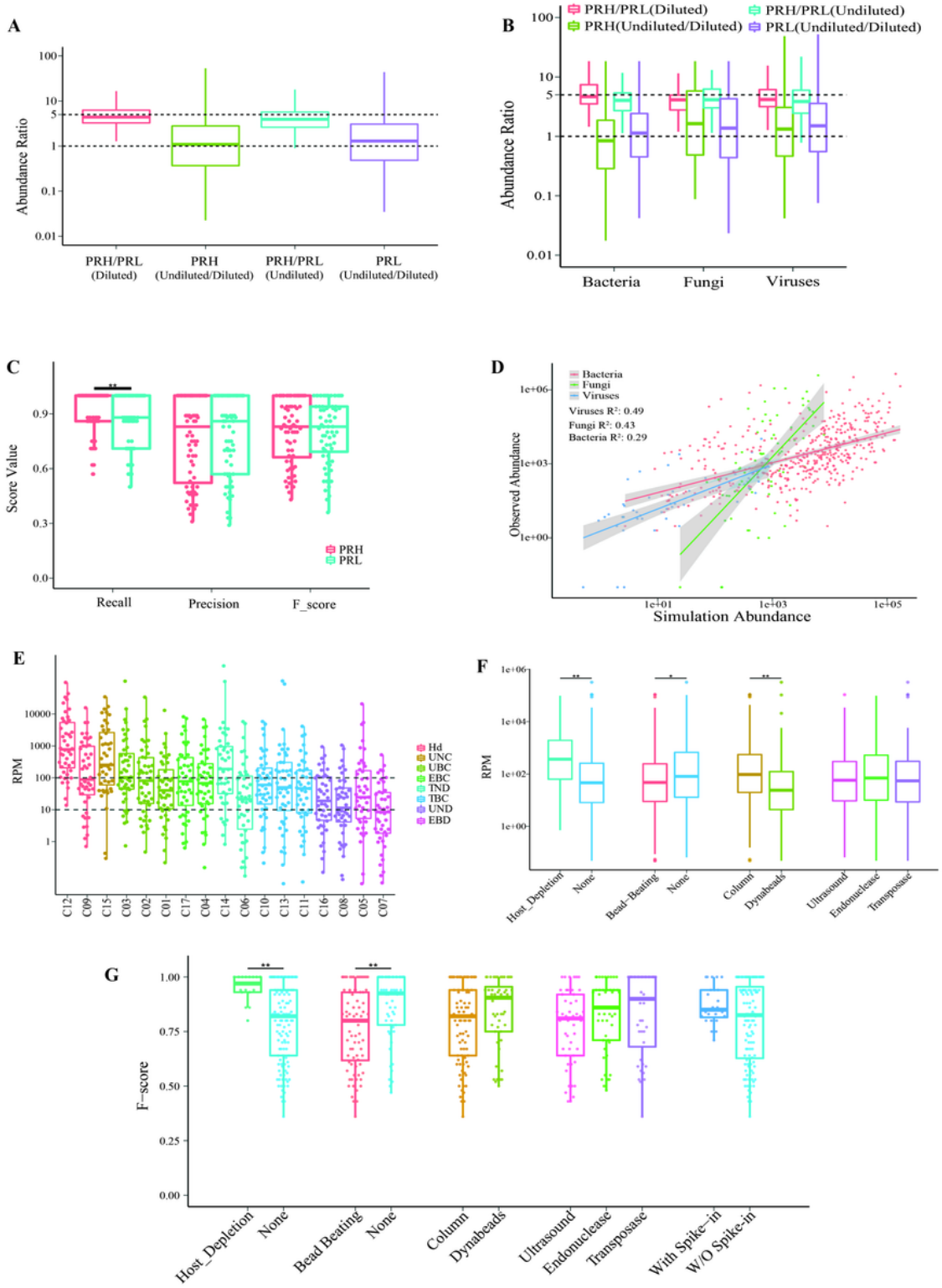

Figure 3

Microbial abundance and pathogen detection by metagenomics. (A-B) Observed abundance ratios between PRH and their PRL counterparts, and undiluted samples and their diluted counterparts, analyzed

together (A) or by microbial type (B). (C) Correlations between relative abundance and performance. (D) Correlations between observed and expected abundances in three microbial types, using the Reads per million (RPM) of HPV virus within each sample as internal control for normalization. (E-F) RPM of microbial detection varied across sites grouped by workflow, Hd: Host-deplete, UNC: Ultrasound, None, Column; UBC: Ultrasound, Bead–beating, Column; EBC: Endonuclease, Bead–beating, Column; TND: Transposase, None, Dynabeads; TBC: Transposase, Bead–beating, Column; UND: Ultrasound, None, Dynabeads; EBD: Endonuclease, Bead–beating, Dynabeads. (E) or key technical variables (F). *, $P<0.05$, **, $P<0.01$ by Wilcoxon rank sum test. See also fig. S3-S5. (G) Correlations between F-score and major workflow-dependent technical variables, **, $P<0.01$.

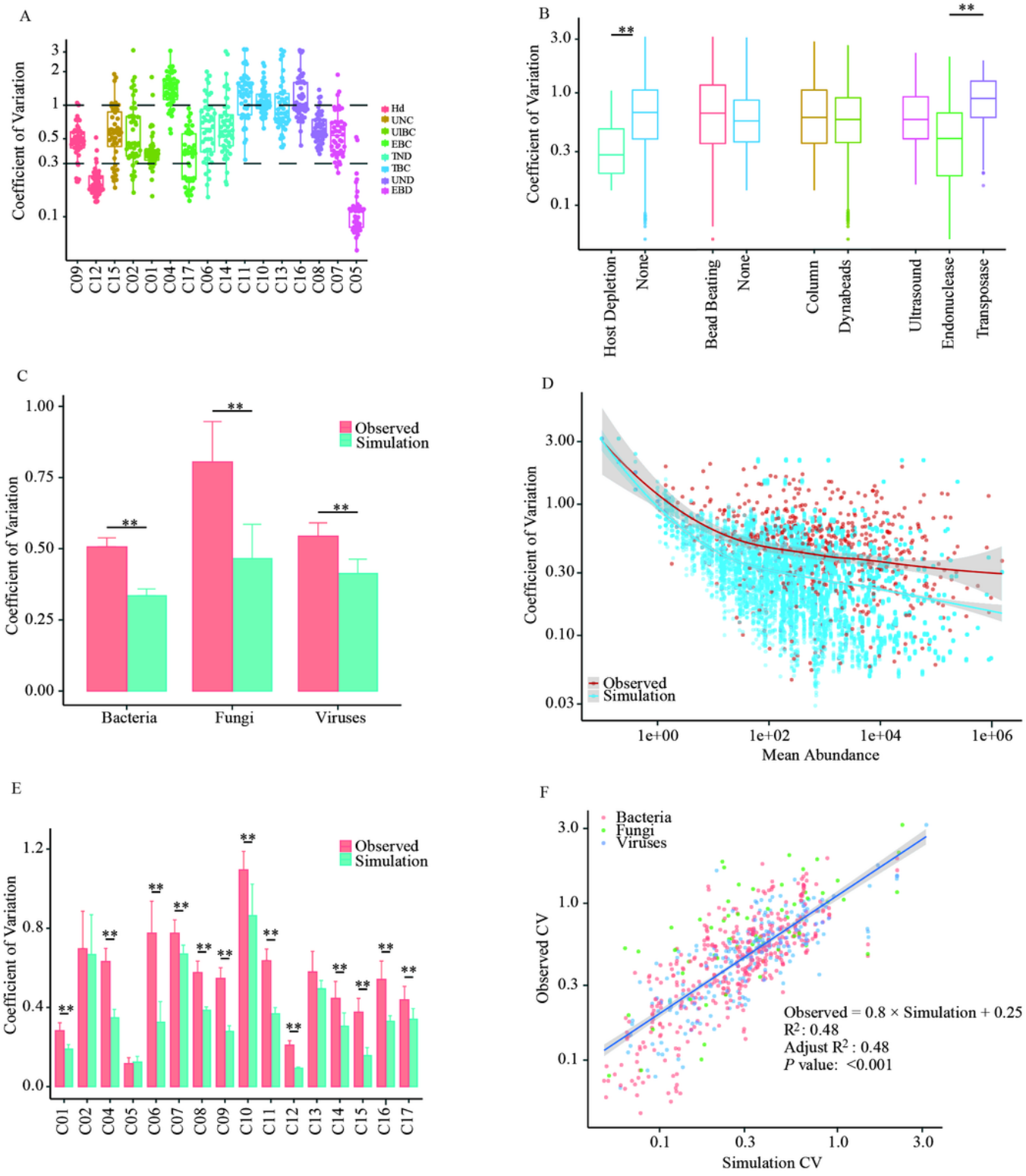

Figure 4

Assessment of assay reproducibility of pathogen metagenomics. (A-C) Coefficient of variance (CV) of microbial mapped reads across sites sorted by workflows (A), grouped by technical variables (B), or microbial type (C). (D-E) Comparison between the overall CV (observed) and the sampling CV (simulated), with data from all sites analyzed together (D) or individually (E). (F) Linear regression with sampling simulated CV as the independent variable. *, P<0.01 by Wilcoxon rank sum test. See also table S6.

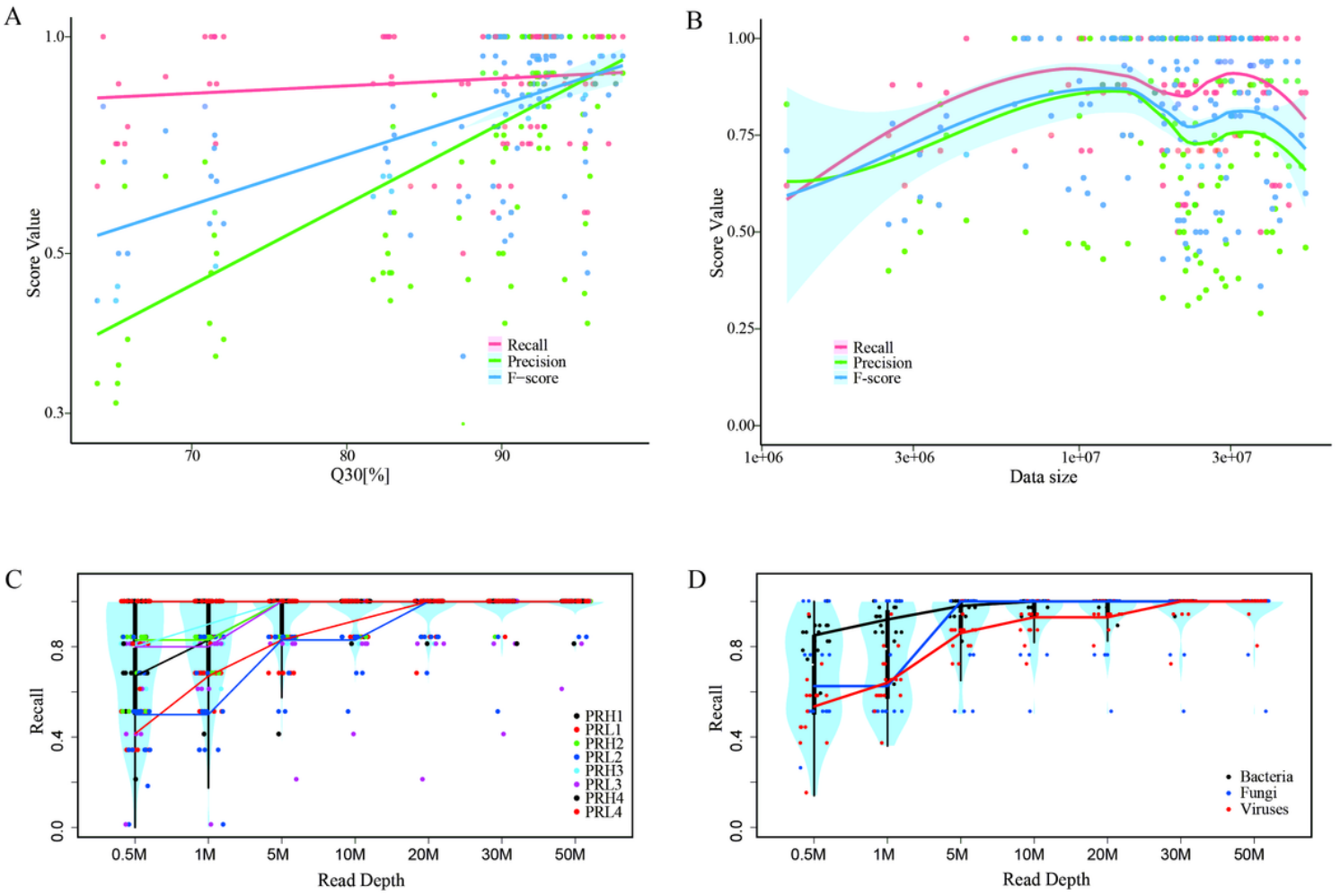

Figure 5

Associations between assay performance and major technical variables. (A) Correlations between performance metrics and sequencing quality as inferred by Q30 score. (B) Correlations between site performance metrics and read depth. (C-D) Performance of pathogen detection as measured by Recall as a function of read depth. Data were grouped and analyzed by each reference reagent (C) or microbial type (D). See also fig. S6-S7.

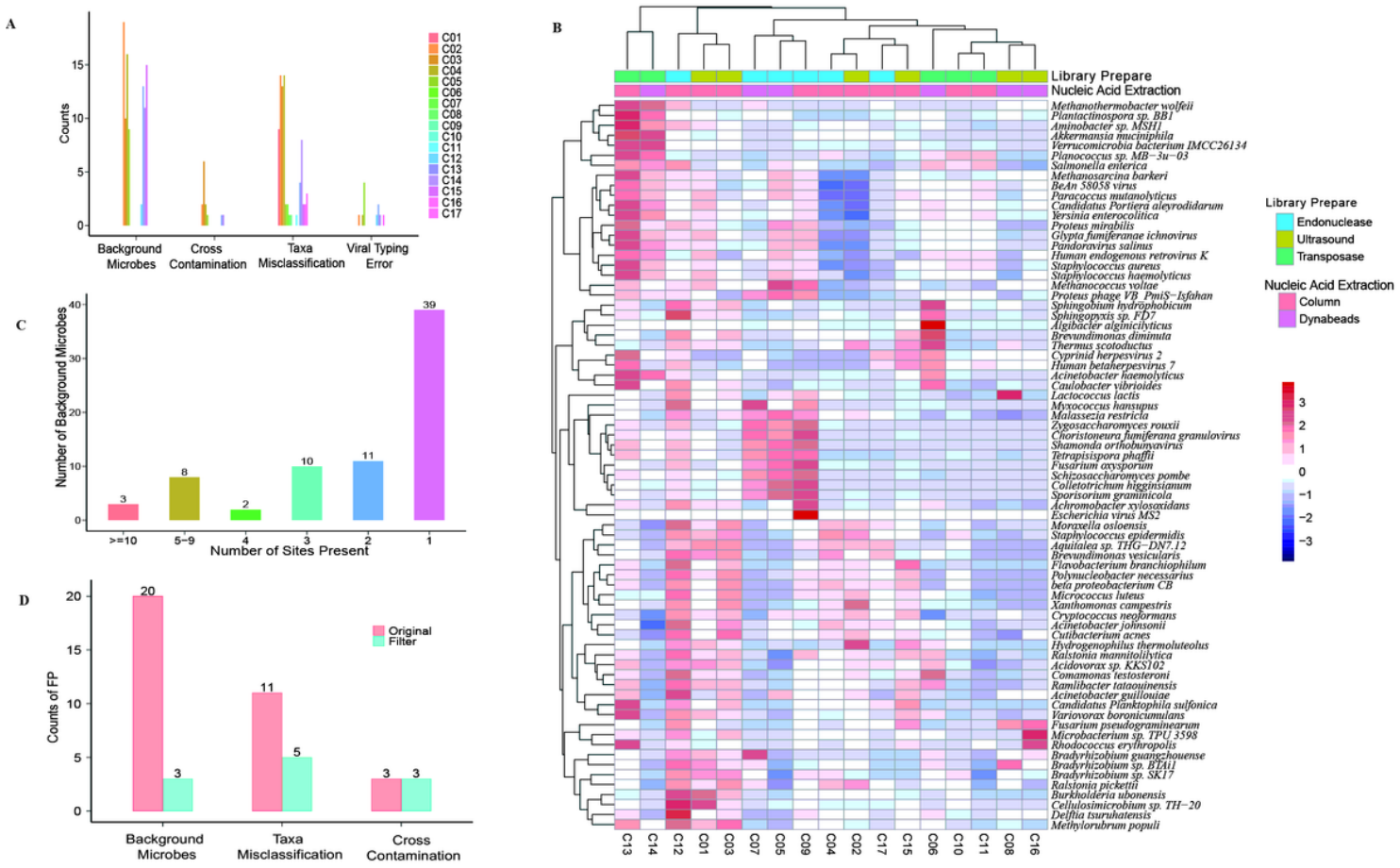

Figure 6

Reducing false positives in pathogen metagenomics. (A) Four main causes of FP. (B) Heatmap showing the top 25 potential background microbes at each site. See also fig. S8-S10. (C) Improved assay specificity after applying FP filters. (D) Summary of the prevalence of background microbes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx