OXFORD (GIGA)^n SCIENCE

TECHNICAL NOTE

IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring

Katrina L. Kalantar [1,*], Tiago Carvalho [1], Charles F.A. de Bourcy [1],
Boris Dimitrov[1], Greg Dingle[1], Rebecca Egger[1], Julie Han[1], Olivia B. Holmes[1],
Yun-Fang Juan[1], Ryan King[1], Andrey Kislyuk[1], Michael F. Lin[1],
Maria Mariano[1], Todd Morse[1], Lucia V. Reynoso[1], David Rissato Cruz[1],
Jonathan Sheu [1], Jennifer Tang[1], James Wang[1], Mark A. Zhang[1],
Emily Zhong[1], Vida Ahyong [2], Sreyngim Lay[3], Sophana Chea[3], Jennifer
A. Bohl [3], Jessica E. Manning [3], Cristina M. Tato [2] and Joseph
L. DeRisi [2]

[1]Chan Zuckerberg Initiative, Science, PO Box 8040 Redwood City, CA 94063, USA; [2]Chan Zuckerberg Biohub,
499 Illinois St, San Francisco, CA 94158, USA and [3]Malaria and Vector Research Laboratory, National Institute
of Allergy and Infectious Diseases, Phnom Penh, Cambodia

*Correspondence address. Katrina L. Kalantar, Chan Zuckerberg Initiative, Science, PO Box 8040 Redwood City, CA 94063, USA. E-mail:
katrina.kalantar@chanzuckerberg.com http://orcid.org/0000-0003-0281-4625

Abstract

Background: Metagenomic next-generation sequencing (mNGS) has enabled the rapid, unbiased detection and identification of microbes without pathogen-specific reagents, culturing, or *a priori* knowledge of the microbial landscape. mNGS data analysis requires a series of computationally intensive processing steps to accurately determine the microbial composition of a sample. Existing mNGS data analysis tools typically require bioinformatics expertise and access to local server-class hardware resources. For many research laboratories, this presents an obstacle, especially in resource-limited environments. **Findings:** We present IDseq, an open source cloud-based metagenomics pipeline and service for global pathogen detection and monitoring (https://idseq.net). The IDseq Portal accepts raw mNGS data, performs host and quality filtration steps, then executes an assembly-based alignment pipeline, which results in the assignment of reads and contigs to taxonomic categories. The taxonomic relative abundances are reported and visualized in an easy-to-use web application to facilitate data interpretation and hypothesis generation. Furthermore, IDseq supports environmental background model generation and automatic internal spike-in control recognition, providing statistics that are critical for data interpretation. IDseq was designed with the specific intent of detecting novel pathogens. Here, we benchmark novel virus detection capability using both synthetically evolved viral sequences and real-world samples, including IDseq analysis of a

nasopharyngeal swab sample acquired and processed locally in Cambodia from a tourist from Wuhan, China, infected with the recently emergent SARS-CoV-2. **Conclusion:** The IDseq Portal reduces the barrier to entry for mNGS data analysis and enables bench scientists, clinicians, and bioinformaticians to gain insight from mNGS datasets for both known and novel pathogens.

Keywords: metagenomics; pathogen detection; virus; cloud-based; COVID-2019

Background

Infectious diseases remain a leading cause of morbidity and mortality worldwide. Despite significant advancement in our understanding of infectious disease biology, existing microbiological tests often fail to identify etiologic pathogens in cases of suspected infection. This can be due to a number of causes—failure to isolate an appropriate sample type, preemptive antibiotic exposure precluding growth in culture, lack of suspicion of a particular infection precluding the ordering of an appropriate test, or lack of available specific diagnostic tests due, in part, to limited knowledge of circulating pathogens. This is compounded further by the fact that novel, previously uncharacterized pathogens may also be present. This fact was illustrated vividly by the recent emergence of COVID-19 in Wuhan, China, in early December 2019. Metagenomic next-generation sequencing (mNGS) of nucleic acid from biological samples offers the potential for a universal pathogen detection method, including the detection of novel species. mNGS has great potential as a broad-spectrum surveillance or patient-monitoring tool, especially in low- and middle-income countries where the infectious disease burden remains high [1]. While the expense of sequencing continues to decrease, the challenge of mNGS data analysis, the lack of bioinformatics expertise, and access to sufficient computational resources and storage remains a major obstacle.

mNGS experiments result in millions of sequencing reads generated from the nucleic acid present within a biological sample, which may include complex microbial populations. A primary goal of mNGS data analysis is to determine what nucleic acid derives from the host (e.g., a patient) and what cannot be attributed to the host or environmental contaminants. Further analysis of the non-host sequence may then attempt to determine the relative abundances of different taxa present in a particular sample because this may provide insight into the presence and relevance of potentially pathogenic microbes. This is typically done via alignment of sequencing reads to a reference database. In the context of infectious diseases, identification of pathogens via this approach obviates the need for pathogen-specific reagents or the ability to culture the microbe. This is especially important for microbes that are difficult or impossible to culture, including many viruses, fungal species, eukaryotic parasites, and bacteria [2]. Additional downstream analysis may then be used to elucidate trends in the abundances and relatedness of organisms across samples.

There are several tools available for estimating relative abundance of microbial populations from mNGS data [3–20]. However, running these tools requires bioinformatics expertise and fluency with command line tools. Additionally, pathogen detection in the context of a host organism presents unique informatics challenges beyond microbial abundance estimation. As noted, a substantial fraction of the sample may consist of host sequences that are secondary to the goal of pathogen detection [21]. Existing tools do not perform sensitive removal of host sequences or quality control (QC) steps, thus necessitating the use of separate QC and alignment tools, and therefore additional computational experience in pipelining. A number of tools exist to incorporate multiple pipeline steps alongside reporting capabilities, including OneCodex [22], Sunbeam [23], and SURPI [24]. However, these tools require paid subscription or significant computational resources to build the underlying databases and run the analyses. Consequently, existing tools are not sufficient to support new applications of mNGS in poorly resourced settings where the detection of infectious agents could have a major effect on population health.

Here, we describe IDseq, an open source cloud-based service for pathogen detection and monitoring. IDseq is a continuously evolving service that enables robust and reproducible analysis of mNGS data for microbial identification, regardless of sample type or host species. We first describe the technical aspects of the IDseq pipeline implementation, including host filtration and QC, assembly-based alignment, and downstream reporting and visualization tools. We then evaluate the performance of the IDseq pipeline, first on a set of standard mNGS benchmark samples as compared to other tools aimed at providing taxonomic abundance estimates from mNGS data, and second on a simulated dataset to evaluate the ability to detect divergent viruses. Finally, we provide two case studies to demonstrate the application of IDseq. First, we apply it to a subset of samples from a previously published report that sought to investigate unknown etiologies of pediatric meningitis [1]. Second, we describe the performance of IDseq in the context of a real-world nasopharyngeal swab, processed and uploaded to IDseq from Phnom Penh, Cambodia, with respect to the emerging viral pathogen SARS-CoV-2. By combining an intuitive web application, a cloud-based pipeline, and downstream visualization tools, IDseq enables investigation of mNGS data for pathogen detection and monitoring, especially suited for researchers with limited computational resources. Importantly, IDseq also enables facile collaboration and data sharing, while enhancing data analysis reproducibility across organizations and countries.

Implementation

IDseq bioinformatics pipeline

The IDseq Portal [25] is a cloud-based, open source bioinformatics platform that enables detection of microbial pathogens from raw next-generation sequencing (NGS) reads. The IDseq pipeline is conceptually based on previously implemented pipelines [1, 26] but is optimized for scalable Amazon Web Services (AWS) cloud deployment (Fig. 1). Here, we describe v3.13 of the IDseq pipeline. Up-to-date pipeline documentation can be found at [27].

The IDseq pipeline ingests raw, short-read sequencing data (either RNA- or DNA-sequencing from any sample type), which can be uploaded from local sources via the web interface or the command line interface (CLI) or directly from Illumina's BaseSpace platform (BaseSpace, RRID:SCR_011881). Sequence analysis proceeds through 3 main phases: (i) host filtering and QC, (ii) assembly-based alignment, and (iii) reporting and visualization (Fig. 1A).
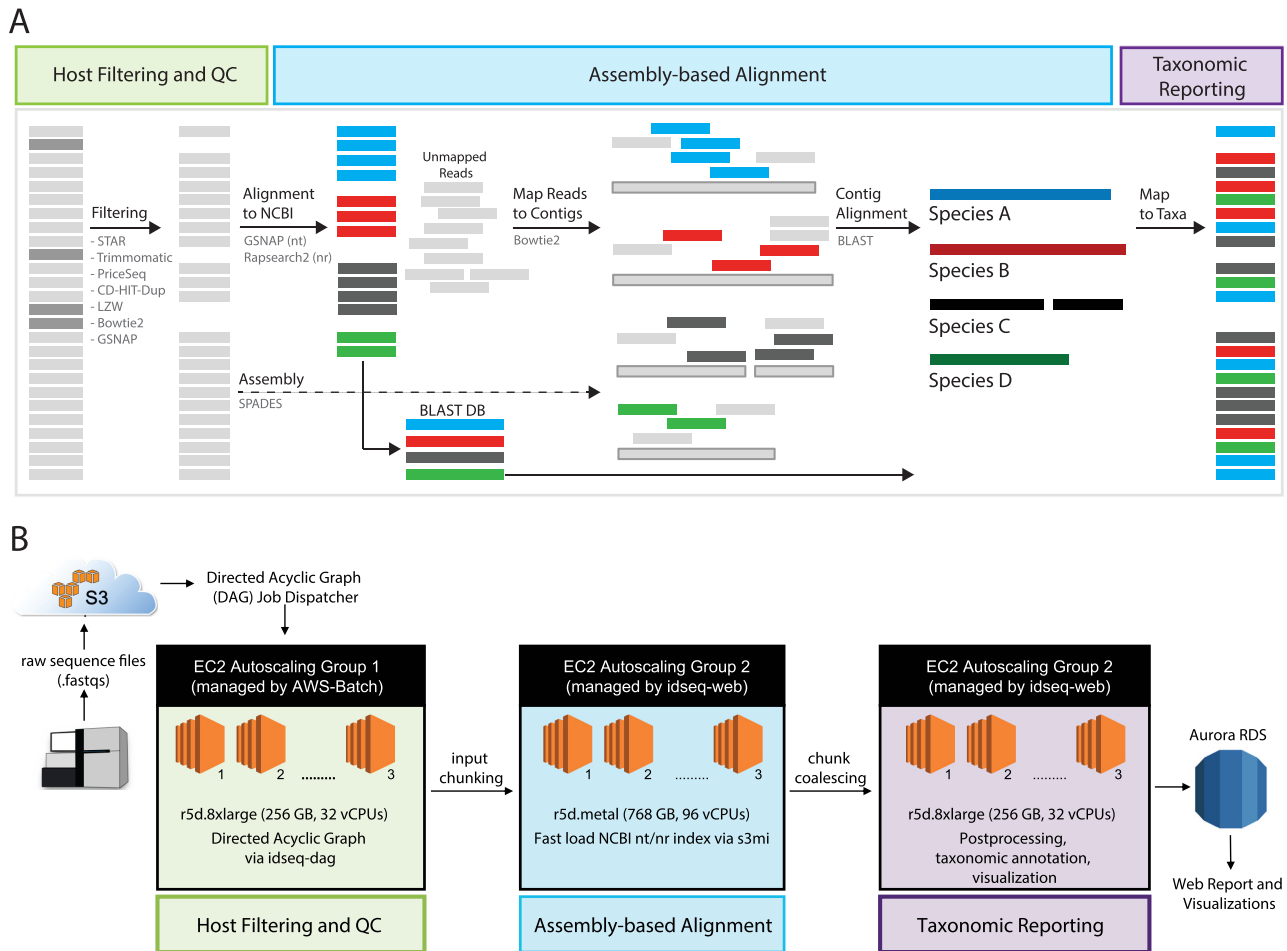
Figure 1: (A) Overview of the IDseq pipeline steps and data analysis workflow. The IDseq pipeline for pathogen discovery is composed of several steps, including host filtering and QC, assembly-based alignment, and taxonomic aggregation and reporting. Each step comprises a number of existing bioinformatics tools. **(B)** The IDseq pipeline is optimized for AWS cloud computational infrastructure. Each of the core pipeline steps (host filtering and QC, assembly-based alignment, and taxonomic aggregation and reporting) is managed by EC2 Autoscaling Groups.

Host filtering and QC

The first phase of the pipeline begins with validation of input files (single- or paired-end .fastq or .fasta files from short-read sequencing libraries). Currently, raw read files are arbitrarily capped at 150 million reads, a threshold that is, according to our experience, larger than most single metagenomic samples. Most mNGS samples processed for pathogen detection are sampled from a potentially infected host organism, and thus the majority of sequencing reads derive from the host organism itself [21]. IDseq performs *a priori* subtraction of host sequences via STAR (Spliced Transcripts Alignment to a Reference) alignment of raw reads to a host-specific database (STAR, RRID:SCR_015899) [28]. IDseq is host-agnostic and allows researchers to select from several available hosts including human, mouse, pig, ticks, and mosquito, among others. For example, human host samples are aligned to the HG38 reference database (GCA_000001405.15), while mosquito samples are aligned to a combined collection of reference genomes from *Culex* and *Aedes* species as well as other diptera. Reads that align to the selected host genome are removed from the analysis. For hosts with well-annotated genomes, individual gene counts may be saved for offline transcriptome analysis, provided appropriate consent in the case of human subject research. Such host-based analyses have been

shown to complement metagenomic analysis for pathogen detection [29]. For all host organisms, sequences for optional spike-in RNA controls developed by the External RNA Controls Consortium (ERCCs) are automatically recognized for downstream steps.

Next, IDseq performs a series of QC steps, as outlined in Fig. 1. First, Trimmomatic [30] trims Illumina adapters. Low-quality reads, duplicates, and low-complexity reads are then removed using the Paired-Read Iterative Contig Extension (PRICE) computational package (PRICE, RRID:SCR_013063) [31], the CD-HIT-DUP tool v4.6.8 (CD-HIT, RRID:SCR_007105) [32], and a filter based on the Lempel-Ziv-Welch (LZW) compression score, respectively. Regardless of the host genome, the data are scoured to remove all remaining human sequences using Bowtie2 against the HG38 reference database [33] and gmap-gsnap against a more stringent database including sequences combining both HG38 and chimpanzees (*Pan troglodytes*) [34]. This step is especially important in the case of vector research, where blood meals may contain human sequences. At each step, the total number of reads remaining in the analysis is computed and these basic QC metrics (including total non-host reads, percent passing QC, and duplicate compression ratio) are provided both in the user interface, as well as via download.

While the host-filtering and QC steps performed by the IDseq pipeline serve primarily to reduce the computational burden and noise in downstream alignment steps, these metrics can also provide a resource for evaluating and troubleshooting sample preparation steps. The proportion of reads lost at each step may provide insight into possible sample degradation, fragment size, sequencing quality, or library complexity. IDseq's automatic estimation of ERCC abundances enables back-calculation of the total input nucleic acid content and estimation of the lower limit of detection and increases the ability to distinguish contaminants [35]. ERCC spike-ins are increasingly recognized as a best practice for addressing the challenges associated with distinguishing background contamination from true microbial populations (Methods) [36].

Assembly-based alignment

To assign taxonomic identities to each read, an assembly-based alignment procedure is used. First, filtered short-read sequences are aligned to the NCBI nucleotide (nt) and non-redundant protein (nr) databases [37] using GSNAPL [34] and RAPsearch2 [38], respectively (Fig. 1A). GSNAPL is a specialized instance of the gmap-gsnap package written by Tom Wu, intended for very large genome databases.

The NCBI database indices are updated biannually, or as needed, via direct pull from NCBI. The index version is tracked for each pipeline run, providing for versioned results. Putative accessions are assigned to each read using the NCBI accession2taxid database [39], and a BLAST+ (v2.6.0) [40] database is constructed on the fly from the set of putative accessions (1 database for each, nt and nr). In parallel, short reads are *de novo* assembled into contigs using SPAdes (SPAdes, RRID:SCR_000131) [41]. Raw reads are mapped back to the resulting contigs using Bowtie2 to identify the contig to which each raw read belongs. Finally, each contig is aligned to the set of possible accessions represented by the BLAST database generated in the previous step, thereby improving the specificity of alignments to all the underlying reads, especially for homologous regions where short reads may align equally well to multiple different accessions.

Reporting and visualization

Where alignments exist, taxonomic identifiers (taxID) for each of nt and nr are assigned to each read. If there exist alignments with equivalent scores to multiple species' taxIDs, then a single taxID is selected at random. If a read was incorporated into a contig, it is assigned the taxID belonging to the NCBI accession to which its parent contig was assigned, as described above. If the read does not assemble into a contig, it is assigned the taxID of the NCBI nt and nr accessions to which it mapped in the initial short-read alignment phase. The results are then aggregated to produce NT and NR counts for each taxID at both the species and genus level. Reads matching GenBank records in the superphylum Deuterostomia are removed, given the high likelihood that such residual reads are of host origin.

The IDseq Portal provides a number of different methods for interpretation of the pipeline results (Fig. 2). First, relevant QC metrics and pipeline run information, including the number of reads remaining at each step of the host and quality filtering steps, as well as estimates of internal control abundances, are provided for each sample (Fig. 2A and B, Methods). The single-sample report tables provide key metrics for each taxon identified in the sample, including the total number of reads aligning to the taxon (in both NT and NR) as well as contig statistics from the assembly-based alignment step (Fig. 2C). The tree view enables rapid assessment of taxonomic related-

ness of microbes identified in the sample (Fig. 2D). For all views of the data, a wide range of user-selectable compound query and filtering tools are made available, enabling facile investigation of the data. For each taxonomic category, IDseq also provides 1-click downloads of the corresponding underlying reads and contigs. Furthermore, coverage plots for contigs relative to all corresponding accessions to which they map are automatically generated (Fig. 2F). To assist with distinguishing microbial signal from reagent and environmental contamination, IDseq supports background model generation, which allows researchers to evaluate the significance (reported in z-scores) of relative abundance estimates for taxons in samples of interest as compared to water-only or other environmental control sample collections. Altogether, the single-sample report and associated filtering functionality enables evaluation of taxonomic hits. More documentation on specific metrics can be found at https://help.idseq.net [27].

To facilitate visualization and hypothesis generation across multiple samples, the IDseq portal provides user-customizable taxon heat maps (Fig. 2E). For advanced users, the pipeline visualization tool clearly documents the input parameters at each step of the analysis pipeline and provides download access to the input and output files at each step so data can be made available for offline analysis (Supplementary Fig. S1), such as phylogenetics.

AWS cloud infrastructure

IDseq is optimized for scalable Amazon Web Services (AWS) cloud deployment (Fig. 1B). Bioinformatics data processing jobs are orchestrated by the IDseq pipeline directed acyclic graph (DAG [43]) and carried out on demand as Docker containers using AWS Batch. Alignments to the NCBI database are executed on dedicated auto scaling groups (ASG) of Amazon Elastic Compute Cloud (EC2) instances, with the number of server instances varied with job load. Fast downloads of the NCBI database from the Amazon Simple Storage Service to each new server instance are enabled by the open source tool s3mi [44].

Versioning and development

IDseq is an open source software tool under continued development across two GitHub repositories—one that hosts the web interface [45] and one that hosts the pipeline code [43]. Modifications to the web interface, which are deployed twice-weekly, do not affect the analysis results. To provide a record of features and how to use them, full documentation can be found at https://help.idseq.net [27].

Updates to the pipeline code may affect analysis results. Therefore, IDseq has adopted a semantic versioning system. Changes implemented to each version are listed in the README file. For each pipeline run, the pipeline and NCBI database versions are also tracked. Major changes to the pipeline outputs result in a major version number update (2.x to 3.x) and are communicated broadly to researchers via email updates. The change from IDseq v2.x to v3.x involved the incorporation of the current assembly steps to refine alignment results, which improved the ability to resolve taxonomic identities in potentially homologous regions. Small changes to the pipeline that may still affect downstream results are indicated by an increased minor version number. For example, addition of a minimum alignment length filter to improve specificity of NT alignments caused a version change from 3.9.4 to 3.10.0. Changes to the pipeline that do not affect the
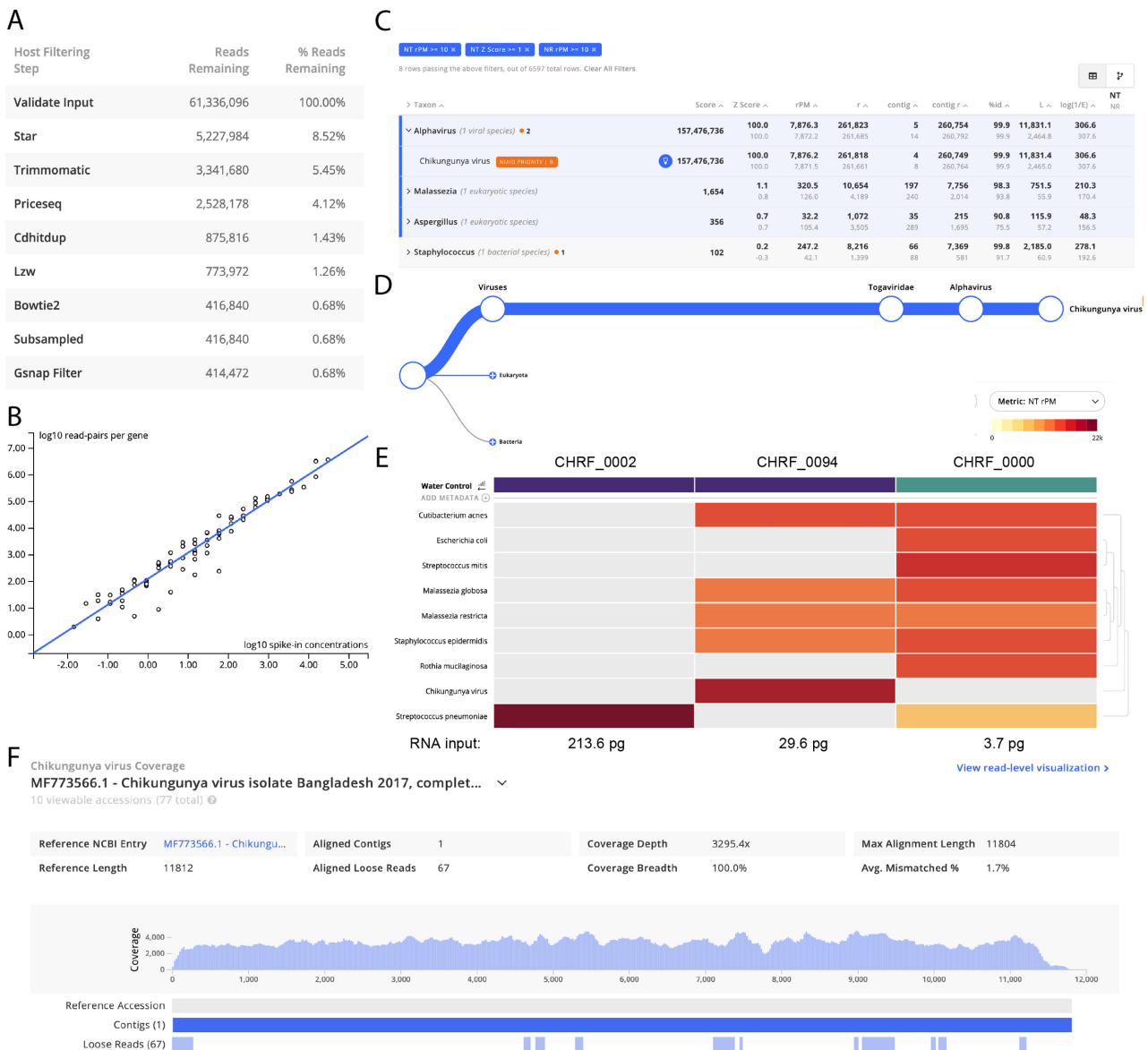
Figure 2: The IDseq web application provides multiple easy-to-use visualizations to help the user assess the quality and content of their sample. Screenshots taken from the IDseq Portal correspond to the re-analysis of samples from a study of etiologies of pediatric meningitis originally published by Saha et al. [1] (see section Application I). CHRF_0002 and CHRF_0094 are CSF samples from pediatric patients with meningitis due to *Streptococcus pneumonia* and chikungunya virus, respectively. CHRF_0000 is a water control. (**A**) Table of reads remaining during each step of the host filtration step (for CHRF_0094)—interpretation of the relative loss at each step in can provide insight into the quality of the library preparation and sequencing run. (**B**) Automatic quantification of ERCC counts from sample CHRF_0094; ERCC quantification enables back-calculation of input RNA concentration. (**C**) The results for a single sample (CHRF_0094) are presented as a table, with key metrics for interpreting taxon alignment quality. (**D**) The tree view indicates the relative abundance of sequences and their taxonomic relationship within a particular sample; shown is the relative abundance of chikungunya virus reads in CHRF_0094. (**E**) The results from multiple samples can be compared using the IDseq heat map view, with associated metadata (purple = CSF, blue = water control). The interactive heat map visualization can be viewed at [42]. The heat map is especially powerful when analyzing trends across a larger number of samples. (F) Coverage of chikungunya virus in CHRF_0094; the coverage visualization enables rapid interrogation of genome coverage.

results are indicated by incremental minor version (i.e., 3.13.1 to 3.13.2).

Continued development on IDseq aims to (i) improve the computational efficiency and accuracy of the results, (ii) expand the integration with other tools to enable researchers' flexibility in the downstream analysis of their processed results, and (iii) support the expanding number of mNGS sequencing platforms

that will be used by researchers for pathogen detection globally. A suite of benchmarking samples are used for analysis of additional pipeline updates as discussed below.

Additional documentation and guides for getting started with IDseq can be found at [27]. The code is open source and available in the GitHub repositories listed in Supplementary Table S1.

Results

Evaluation of IDseq on external benchmark datasets

IDseq analysis of unambiguously mapped datasets

A recent study evaluated the performance of 20 taxonomic classifiers for mNGS data on 10 reference datasets that are commonly used for benchmarking, containing computationally simulated reads from between 12 and 525 bacterial species [46]. It evaluated performance using 2 metrics—the area under the precision recall curve (AUPR) and the L2 distance. The AUPR evaluates the ability to detect the presence of microbes (binary presence/absence) above a relative abundance threshold, taking into consideration the precision and recall rates as said threshold is adjusted. A species-level AUPR of 1.0 indicates that there is a threshold (proportion of reads) above which all true-positive species can be identified with no false-positive species. The L2 distance provides a complementary metric that considers the similarity in relative abundances between the results and the ground truth.

We evaluated the performance of the IDseq pipeline on these same datasets (Methods, Supplementary Table S2). Samples took a mean of 3 hours (minimum = 1.6 hours, maximum = 10 hours) to process on IDseq pipeline version 3.13, with the NCBI database version from September 2019. Performance metrics (AUPR and L2 distance) were computed separately for the NCBI nt and nr results and compared to those published recently by Ye et al. (idseq_nt and idseq_nr, Fig. 3) [46]. IDseq provides an automated pipeline but at the cost of inability to easily swap in new databases. Therefore, we compared our results against those reported by Ye et al. for the "default database" of other tools. The performance metrics may inherit biases due to differences in the reference database contents, as well as recency of input sequences.

Deep dive of unambiguously mapped dataset results

The IDseq pipeline demonstrated performance comparable to that of the other mNGS tools tested (Fig. 3). The unambiguously mapped datasets demonstrated limited resolution for distinguishing the tools when evaluated by AUPR and L2 because most tools show relatively high performance (with AUPR scores >0.8 at the species level, Fig. 3A). Consistent with Ye et al. [46], we observed that the greatest differences between tools were in the reduced precision at high recall. IDseq protein alignments (NR) demonstrated greater AUPR than IDseq nucleotide (NT) across most datasets but consistently identified more taxa at low abundance (<1%), therefore resulting in reduced precision (Fig. 3C). Meanwhile, IDseq NT exhibited increased specificity. IDseq NT and NR had a mean AUPR across all the datasets of 0.9627 and 0.9633, respectively. The top mean AUPR of any single tool was achieved by metaothello (0.9661), followed by Kraken2 (0.9635). Given Kraken2's performance on the unambiguous benchmark datasets and its wide adoption for relative microbial abundance estimation, additional analyses focused on comparison against Kraken2 (Fig. 3B and C). Another distinguishing factor between the tools was in the number of reads that were "unclassified" across multiple datasets. mmseq2, metaothello, kaiju, and bracken consistently left >10% of reads unclassified. IDseq (NT and NR) removed a mean of 10% of the reads during host filtering and QC steps, but of the remaining sequences, a mean of <1% of reads were unmapped across the 10 datasets. This can be attributed in part to IDseq always assigning reads to a species when an alignment exists (increasing sensitivity at the expense of specificity) and secondarily to the use of assembled contigs to refine alignments where short reads may have been unmapped.

To further investigate differences between the tools, we evaluated the results for each dataset independently (Fig. 3B). IDseq NR demonstrated lower precision across all datasets than many other tools, including IDseq NT and Kraken2 (Fig. 3C, Supplementary Fig. S2). The ATCC Staggered dataset, which includes several microbes present at very low abundance, yields the lowest AUPR of all samples tested via IDseq NR, consistent with findings in Ye et al. [46] that protein-based classifiers consistently struggled to identify the low-abundance taxa amongst other low-abundance false-positive results. Meanwhile, IDseq NT demonstrated reduced performance on the NYCSM dataset (Supplemental Text). IDseq's use of the full NCBI nt and nr databases resulted in relatively high performance for the Buccal dataset. Ye et al. discuss that the Buccal dataset was a low-performing outlier for most evaluated classifiers due to inclusion of reads from a species with only contig-quality reference, which is not included in most default databases.

The IDseq web portal is designed to provide researchers with the choice of using either NT or NR results, or both in conjunction with each other. For example, the impact of spurious NR alignments can be mitigated by requiring a corresponding alignment with IDseq NT. Using this strategy, the performance of IDseq was evaluated, considering the NT relative abundances reported for taxa with both NT $r > 0$ and NR $r > 0$ (idseq_ntnr, Fig. 3, Supplementary Fig. S2). We observed that requiring concordance resulted in the greatest mean AUPR across all other tested tools (0.9673) and increased the precision of IDseq above that of either NT or NR alone.

Altogether, these results highlight some key trade-offs with respect to relative abundance estimation of bacterial species. IDseq is capable of identifying organisms with respect to the latest versions of NCBI and demonstrates relatively high recall (Fig. 3D). But use of the full NCBI database may result in false-positive alignments at low abundance, which can reduce precision (Fig. 3C). This is especially true for bacterial taxa, with homology in the 16s ribosomal RNA (rRNA) regions. In the context of pathogen identification, it has been observed that infecting agents often comprise the majority of sequencing reads [29]. For such datasets, the reduced precision for abundance estimation at low levels has less of an effect. One case in which this may not be true is that of viral respiratory infection, whereby a small number of sequencing reads may be indicative of infection. In this case, targeted analysis using IDseq to filter for only viral reads will improve sensitivity. Meanwhile, researchers interested in evaluating highly complex microbiome composition at the species and strain level may need to bring in other tools to supplement their analyses [47–49] or rely on genus-level estimates provided by IDseq.

Evaluation of IDseq on internal benchmark datasets

To address the gaps between the existing benchmark datasets and the IDseq pipeline's primary use-case for pathogen detection, we tested IDseq's performance on 3 additional datasets specifically designed to evaluate detection of divergent viruses (Methods) and common clinical microbes (Supplemental Text). For each dataset, we evaluated the performance of IDseq (NT and NR), as compared to Kraken2 [15], using per-species recall.

Detection of divergent and novel viruses

Viruses are known to evolve rapidly, and therefore their sequences may diverge from sequences in the known NCBI
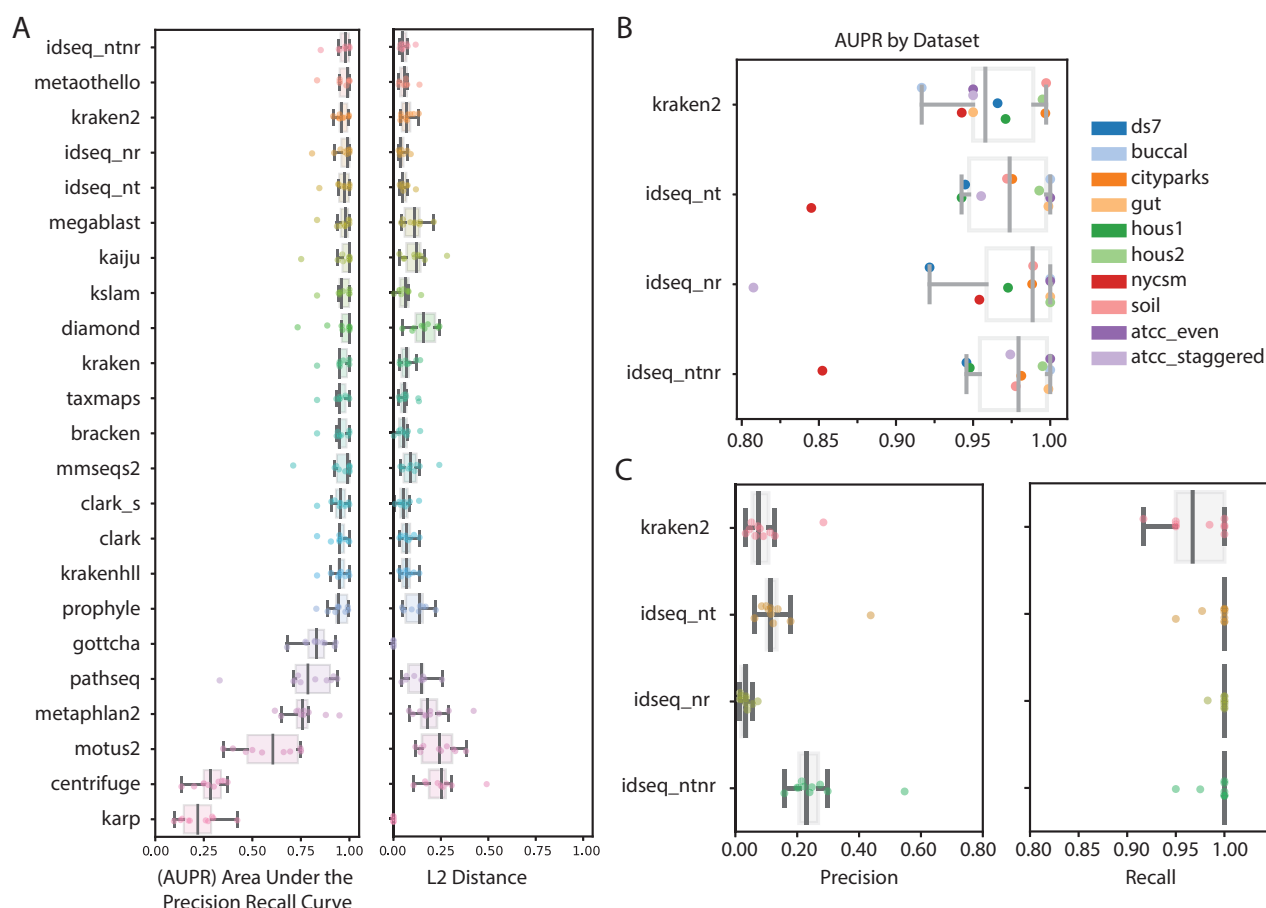
Figure 3: Performance metrics calculated for IDseq (NT and NR), as compared to the values recently published by Ye et al. [46]. (**A**) Area under the precision recall curve (AUPR) and L2 distance values for 22 tools, as evaluated against their default databases. (**B**) The AUPR values for specific benchmark datasets evaluated for 3 tools (Kraken2, IDseq NT, and IDseq NR), including metrics obtained when evaluating basic threshold filters integrating both IDseq NT and NR (idseq_ntnr). (**C**) The precision and recall of the same 3 tools for detecting known taxa. In all boxplots, the median is shown as a dark grey line, with light grey boxes corresponding to the first and third quartiles. Whiskers extend to the farthest data points that are not outliers.

database over relatively short timescales [50]. Maintaining the ability to detect divergent viruses is of paramount concern, given their role in numerous recent outbreaks, including the recent emergence of SARS-CoV-2, the coronavirus responsible for the COVID-19 outbreak [51–54]. The idseq-bench tool was used to generate 17 simulated NGS samples from Rhinovirus C genomes at varying levels of divergence (after *in silico* forward evolution from a reference sequence obtained from the NCBI database), ranging from 100% identical to the reference sequence to 25% similar (at the nucleotide level) (Methods, Fig. 4A, Supplementary Table S3). The resulting samples were uploaded to IDseq (Project HRhinoC Simulation). Meanwhile, the same raw .fastq files (prior to host filtering) were analyzed using Kraken2 (Methods).

Both IDseq NT and Kraken2 identified reads aligning to Rhinovirus C down to 75% sequence divergence (Fig. 4B). Meanwhile, IDseq NR recalled Rhinovirus C alignments down to 70% sequence divergence, demonstrating a greater sensitivity for divergent virus detection. We note that IDseq NR experienced a rapid decrease in total recall (8,558 reads correctly mapping to Rhinovirus C, of 10,000 total and 8,558 passing QC steps at 70% sequence similarity vs 0 reads detected at 65% sequence similarity). This highlights an artifact of the computational cost-saving mechanisms employed by IDseq—whereby a BLAST database

is constructed from only the subset of accessions identified in the initial short-read GSNAP and Rapsearch2 alignments to the NCBI database. In cases where the highly divergent short-read sequences do not match to NT or NR in the initial alignment, the BLAST database will be empty and none of the reads or contigs will map. However, IDseq does provide the ability to download all assembled contigs, enabling offline interrogation of this divergent "dark matter." Manual BLASTx of contigs assembled by SPADES in IDseq to the full NCBI database was able to recover the Rhinovirus C identity down to 55% sequence identity. Future iterations of the IDseq pipeline may aim to automate the manual follow-up steps for divergent viral contigs, as well as incorporating other tools for dark matter investigation to probe for pathogen motifs.

Further comparison of the IDseq (NT and NR) results to Kraken2 shows that Kraken2 initially recovered more of the simulated reads than IDseq (9,964 of 10,000 vs 8,582 for both IDseq NT and NR). This is explained by the QC steps in the IDseq pipeline, which removed ~15% of reads at the PriceSeq filtering step owing to low quality—an expected outcome given that the simulated reads mimic error models of Illumina sequencers (Methods). Of the reads remaining after host filtering, IDseq identified 100% as aligning to Rhinovirus C. This pattern persists down to 95% sequence similarity, at which point
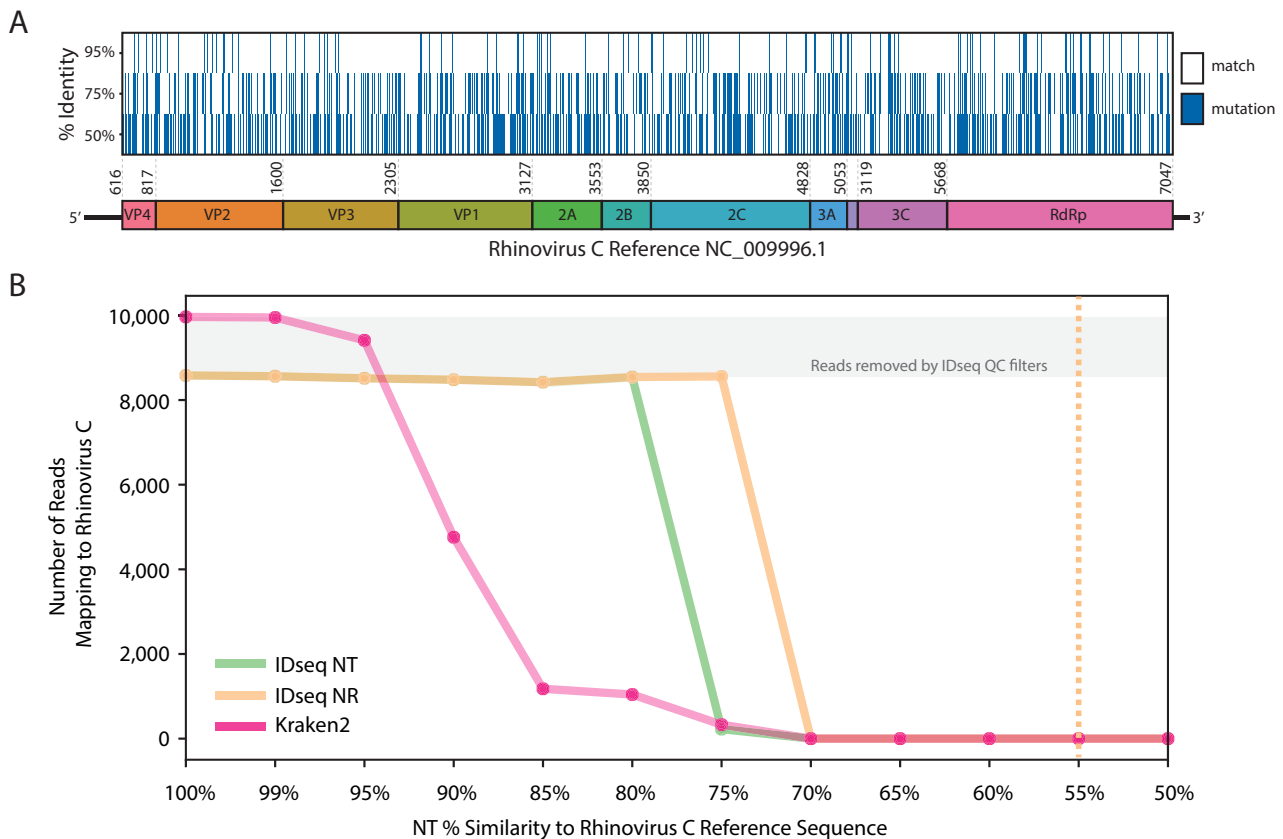
Figure 4: (**A**) Graphic representation of genomic similarity for simulated divergent Rhinovirus C genomes, at 95%, 75%, and 50% similarity to reference sequence NC_0 09996.1. (**B**) Performance of IDseq (NT and NR) as compared to Kraken2 for recovery of reads from simulated divergent Rhinovirus C genomes at varying levels of divergence. The dotted line indicates the theoretical limit for detection of Rhinovirus C achieved by manual BLASTx of IDseq-produced contigs.

Kraken2 begins to identify fewer reads. While some Rhinovirus C reads are identified down to 75% sequence similarity (same as IDseq NT), IDseq NT identified a significantly greater number of reads mapping to Rhinovirus C at increasing levels of divergence. Specifically, at 80% divergence, 8,544 reads were mapped by IDseq NT while only 1,042 reads were mapped by Kraken2. Altogether, these benchmark results are consistent with existing reports of the utility of IDseq NR in detecting divergent viruses [55] and are within the ranges of nucleotide divergence associated with emerging human pathogens (Supplemental Text).

Application I. IDseq for pathogen discovery in cases of pediatric meningitis

The IDseq pipeline is sample-type agnostic, allowing researchers interested in a broad range of scientific questions across a diverse array of host organisms (e.g., humans, mice, mosquitos, ticks, plants, environmental) to obtain relevant microbial information from any sample type (e.g., blood, cerebrospinal fluid [CSF], respiratory fluids, tissue) [1, 56–58]. There are many challenges for data interpretation that are common across mNGS applications, such as impact of PCR amplification on samples with low amounts of input RNA, background contamination, and genomic similarity between short regions of related organisms. Here, through a re-analysis of the IDseq results for 3 CSF samples from a recent study investigating etiologies of pediatric meningitis in Bangladesh [1], we highlight specific IDseq features to address these challenges. The original

study, conducted by Saha et al., included 91 CSF samples (36 positive, 30 negative, and 25 idiopathic) and 6 water controls, processed on IDseq v3.1. We focus on 1 known infection (*Streptococcus pneumoniae*, CHRF_0002), 1 idiopathic sample that was later confirmed to have chikungunya virus (CHRF_0094), and 1 water control (CHRF_0000) (Fig. 2). These samples, for demonstrative purposes, were re-run on IDseq v3.13 and are available in IDseq project CHRF RR007 Example. Key pipeline run metrics for these 3 samples are provided in Table 1.

Sample 0094: A case of neuroinvasive chikungunya virus
CHRF_0094 was a pediatric encephalitis case of unknown etiology that was later determined to be a case of neuroinvasive chikungunya virus. Fig. 2A shows the number of reads removed by each host filtration and QC step. One challenge for mNGS-based pathogen detection is that host sequences dominate the mNGS library. Notably, in CHRF_0094, chikungunya virus reads in sample CHRF_0094 represented <1% of the total sequencing reads. However, after IDseq's host filtering and QC steps, it represented 63% of the remaining non-host reads. A second, widely acknowledged challenge for mNGS data interpretation is the presence of environmental contaminants. Best practices suggest including ≥1 water control with every sequencing experiment [35, 59]. To assist with interpretation of results with respect to control samples, IDseq implements a z-score approach (Methods) first described in Wilson et al. [60]. The z-score statistics computed by IDseq indicate the significance of relative abundance estimates in a sample as compared to the user-

Table 1: Key pipeline run metrics for 3 samples

Parameter	CHRF_0094	CHRF_0002	CHRF_0000
Description	Chikungunya viral meningitis	Streptococcus pneumonia meningitis	Water control
Sample type	CSF	CSF	Water
Collection location	Bangladesh	Bangladesh	Bangladesh
Total reads	61,336,096	141,979,356	135,087,088
ERCC reads	28,094,424	14,875,054	130,150,782
STAR	5,227,984	26,802,224	4,675,916
Trimmomatic	3,341,680	23,770,970	3,440,016
PRICE	2,528,178	20,752,710	1,964,846
DCR	2.89	1.39	4.67
RNA input concentration (back-calculated from ERCCs)	29.6 pg	213.6 pg	3.7 pg
No. of rows (NT rpm > 10, NR rpm > 10, NT Z > 1)	8	2	36
NT rpM	7876.2	22,034.0	NA
NR rpM	7871.5	19,743.9	NA
No. of Contigs	4	143	NA
Alignment L (nt)	11,831.4	75,065.3	NA
Mean % identity	99.9	99.2	NA

The host filtering and QC stage of the IDseq pipeline is composed of several individual steps. The proportion of reads lost at each step can provide insight into sample quality and library preparation. Interpretation of these metrics may be valuable for laboratories evaluating new sample storage techniques, library preparation protocols, etc. These 3 samples, provided as an example, can be investigated in the IDseq portal in Project CHRF RR007 Example. CSF: cerebrospinal fluid; DCR: duplicate compression ratio; L: mean alignment length across all reads and contigs mapping to that taxon; No. of rows: total number of species and genus-level rows in the IDseq sample report; NR: results based on NCBI non-redundant protein (nr) database; NT: results based on NCBI nucleotide (nt) database; rPM: reads per million.

selected background controls—which may include water controls or healthy control samples. A z-score threshold can be imposed to remove taxa that are prevalent in the water or healthy controls. In sample CHRF_0094, 8 rows (4 species from 4 genera) were reported with NT reads per million (rPM) > 10, NR rPM > 10, and z-score > 1 (a relatively stringent threshold used to remove many of the low-abundance taxa for first-pass evaluation). A total of 7,876.2 rPM were associated with chikungunya virus, of which many were associated with the 4 contigs aligning to chikungunya virus. By using the IDseq portal coverage visualization, which displays reads and contigs in association with their top matched GenBank accession, we observe that the longest contig, ~11 kb, represented full-genome coverage of the nearest GenBank accession (Fig. 2F).

Sample 0002: A case of known Streptococcus pneumoniae meningitis
In sample CHRF_0002, IDseq associated 1,927,505 reads by NT with the independently verified pathogen *Streptococcus pneumoniae*, of which 98.1% were assembled into 143 contigs (Table 1). The mean alignment length across all contigs and reads was 75,289.8 bp—driven largely by alignment of long contigs. Despite the large number of contigs and long alignment lengths, the GenBank accession with the greatest coverage (1.8 Mb LR216026.1 *S. pneumoniae* strain 2245STDY5775485 genome assembly, chromosome: 1) had 87.3% coverage breadth. This exemplifies a frequently observed pattern (which is even more pronounced in lower-coverage samples)—whereby coverage of larger bacterial genomes is lower than virtual genomes even for samples with a high proportion of mNGS reads associated with a particular microbe. For many cases, low coverage from mNGS data can preclude confident strain identification in bacterial species that may be useful in a clinical context. Furthermore, low coverage of the transcriptome (via RNA mNGS) may produce a large proportion of alignments in conserved rRNA regions, which may be challenging to disambiguate.

Sample 0000: A water control
In sample CHRF_0000, the duplicate compression ratio (DCR) of 4.67 indicates the possibility of over-amplification of low-biomass nucleic acid input (Table 1). This is common for water samples, where low input nucleic acid is expected. The use of ERCC controls in the library preparation of these samples enabled back-calculation of the total input RNA concentration. This sample was determined to have 3.7 pg of total input RNA, while the 2 infected samples (CHRF_0094 and CHRF_0002) had 29.6 and 213.6 pg, respectively. Thus, while the relative abundance values seem comparable to those in the infected sample, they represent significantly smaller quantities of raw nucleic acid (Fig. 2E). In the original study all water and non-infectious controls (which had low white cell counts and therefore little host or pathogen nucleic acid) had input RNA quantities <4 pg, enabling the use of an input nucleic acid threshold for inclusion in downstream analyses. Additionally, the top 4 organisms (by NT rPM) include *Providencia*, *Cutibacterium*, *Streptococcus*, and *Escherichia*—many of which are known environmental contaminants [36, 61, 62]. The 36 total rows (with NT rPM > 10, NR rPM > 10, and NT z-score > 1) are all present at relatively similar and low abundance levels, characteristic of background contaminants [36, 35].

Application II. Real-time detection of novel coronavirus

IDseq is a globally accessible pipeline for mNGS analysis that has been shown through simulation and practice to be effective in identifying novel and divergent viruses. As an additional real-world example of this utility, we provide a vignette from the recent SARS-CoV-2 coronavirus outbreak. On 30 January 2020, a team of researchers from the CNM-NIAID (National Center for Parasitology, Entomology, and Malaria Control—National Institute of Allergy and Infectious Diseases) collaboration in Cambodia obtained a nasopharyngeal swab sample from a patient with PCR-confirmed SARS-CoV-2 infection. The library preparation and sequencing were completed in-country by 1 February

2020 [63]. Analysis of the sample (4.5 million single reads) using IDseq against an NCBI database version from 17 September 2019, which did not contain the known sequences for SARS-CoV-2 that have since been deposited in NCBI, identified 571 reads aligned to the genus *Betacoronavirus*, with a mean amino acid percent identity of 92.3% (by NR). The sample took 14 minutes to analyze end-to-end and the most abundant species was severe acute respiratory syndrome–related coronavirus, with 542 NT reads (22 contigs) and 571 NR reads (24 contigs), representing ∼33% genome coverage. To quantify the IDseq pipeline's recall for SARS-CoV-2 sequences, we built a BLAST database from the 54 sequences associated with SARS-CoV-2, which had been deposited in NCBI between January and 2 February 2020 as a result of widespread efforts by the global science community. By BLASTing all non-host reads from the sample against the known SARS-CoV-2 sequence database, we identified 584 reads mapping to SARS-CoV-2. As compared to this ground-truth value, IDseq demonstrated 97.8% read-level recall. This indicates that for an emerging threat, IDseq was able to successfully provide information on the presence of a pathogen prior to the existence of full reference genomes associated with the organism. This identification was of paramount public health importance given unclear diagnostic accuracy in the beginning pre-pandemic state.

Discussion

We have introduced IDseq, an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. We described the pipeline analysis steps and demonstrated that the IDseq pipeline achieves performance for taxonomic identification and relative abundance estimation comparable to that of other tools in the field. We showed that IDseq is uniquely suited for detection of divergent viruses and has high sensitivity for detecting human pathogens. Finally, we have shown through two case studies how the IDseq portal enables researchers to rapidly generate insights into their samples' quality, microbial content, and cohort trends. We further highlighted its real-time utility by describing how IDseq was used to analyze sequences associated with the emerging coronavirus SARS-Cov-2 prior to deposition of SARS-CoV-2 sequences into public data repositories. The IDseq web portal provides an easy-to-use access point for computationally intensive analysis of mNGS data. Its sample-type–agnostic implementation enables its application for a broad range of research questions related to understanding the distribution of microbes in a sample. The IDseq pipeline has been a key component in recent studies investigating undiagnosed causes of infection and surveying the landscape of circulating pathogens, in both humans and animals [64, 65].

Benchmarking of mNGS tools is a well-recognized challenge within the field [46, 66, 67]. The choices of tools, parameters, databases, and datasets may all influence the conclusions. Our aim in the present study was simply to test performance relative to other tools. We compare IDseq's default database (NCBI nt and nr) against the default databases for all other tools included by Ye et al. [46]. Although it is possible that other tools' performance would improve given a comparably large database, configuring these details requires computational expertise obviated by the readily usable nature of IDseq. IDseq continues to use the full NCBI nt and nr database given their advantages for detecting divergent viruses and incorporating data on novel bacterial pathogens. However, the large database size results in longer run-times and the lack of curation induces the potential

for noise in alignment results due to errant sequence assignment errors upon upload to the NCBI databases. There is ongoing work by many researchers to evaluate curated databases for mNGS analyses, but for now IDseq continues to update its database biannually. To support continued benchmarking of IDseq and empower researchers to test IDseq's performance for their particular applications, we have released the open source idseq-bench tool, which was used to generate the divergent virus dataset and for evaluating the per-read recall results.

Beyond the informatics nuances between tools, IDseq provides clear advantages for researchers new to mNGS and computational data analysis. First, IDseq is designed and maintained by a team of engineers and managed as a software-as-a-service product, where user support is a key component. User support enables researchers to have confidence that they will obtain results in a timely fashion. Second, the tool's user interface provides a series of advantages for users with limited computational expertise by reducing the challenges associated with installation and configuration, as well as providing meaningful metrics for quality control and interpretation. It maintains transparency on individual pipeline steps through documentation [27], the pipeline visualization tool (Supplementary Fig. S2), and availability of downloads from intermediate files. Together, these resources help researchers new to mNGS get started quickly, while also providing tools to enhance skills in computational biology. Third, the pipeline provides assurance of computational reproducibility, which is an increasingly appreciated priority within the scientific community as dataset sizes and analytical complexity increase. Last, the web-based user interface provides an access point for collaboration and networking—enabling researchers to collaborate seamlessly across countries and institutions, thereby building global networks of expertise that can be accessed by those in resource-scarce settings.

Finally, we highlight that IDseq is not a clinical tool and is intended only for research purposes. IDseq aims to be a valuable resource for researchers in the infectious diseases field but does not intend to become clinically validated. While IDseq can yield insights that inform public health policies, laboratory testing priorities, and real-time decisions for confirmatory clinical testing, clinical validation of the pipeline requires locking of the system for adherence to strict guidelines. IDseq will remain under continued development in order to (i) improve the computational efficiency and accuracy of the results, (ii) expand the integration with other tools to enable researchers' flexibility in the downstream analysis of their processed results, and (iii) support the expanding number of mNGS sequencing platforms that will be used by researchers for pathogen detection globally. Some possible future directions for improvements to the IDseq pipeline have been discussed throughout. Notably, IDseq's current assembly-based alignment steps result in failure to automatically identify divergent viruses beyond 70% divergent, while BLASTx of IDseq-generated contigs can enable detection down to 55% divergent. Automating full NCBI BLASTx of putative viral contigs would simplify offline analyses. Similarly, we showed that IDseq NR had reduced precision, which made relative abundance estimation of low-abundance taxa challenging. Allowing for non–species-specific mappings or propagating estimates of species-level ambiguity to increase species-level resolution for low-abundance taxa may provide another avenue for continued development. Finally, continued integration with other analysis tools and sequencing technologies will further enhance the usability of IDseq for mNGS data analysis.

IDseq reduces the need for much of the computational expertise and access to large-scale computing resources that have

traditionally been barriers for conducting mNGS data analysis. The IDseq portal provides an easy-to-use interface that enables researchers around the world to upload samples and generate hypotheses with relevant implications for global health and infectious disease tracking as diseases emerge.

Methods

Raw pipeline commands

The IDseq pipeline uses several publicly available academic bioinformatics tools. The raw commands and parameters used for each step in the pipeline are available for each pipeline version in the pipeline visualization (Supplementary Fig. S1), which can be viewed for any sample in IDseq. Technical documentation is available [68].

Automatic ERCC quantification

The ERCC developed a common set of external RNA controls that can be used to control for a variety of sources of variation on RNA expression attributed to experimental factors (including the quality of the starting material, the level of cellularity and RNA yield, the sequencing platform, and the person performing the experiment). In the context of pathogen detection, mNGS libraries often contain extremely low quantities of RNA input. It has been shown that during library preparation, samples with low input experience amplification background contaminants [35]. ERCC controls can be used to mitigate the effect of low input libraries and to quantify the total input. To enable researchers to rapidly assess the quality of their libraries and the limit of detection, IDseq provides ERCC counts for each sample. During the host-filtering steps, the raw sequencing reads are aligned to the ERCC reference sequences and counts are generated by STAR –genecounts option [28]. These values are then available for download, as well as visualized in the user interface (Fig. 2B).

IDseq z-score and aggregate score metrics

Given the sensitivity of mNGS, it is common to identify contaminating microbial sequences derived from laboratory contaminants, reagents, collection tubes, etc. There exist numerous approaches to assist in distinguishing background contaminants from true microbes [35, 60, 69]. IDseq implements a previously described z-score method for background correction [60]. Researchers can create a background model by selecting control samples sequenced via their standard laboratory protocols or select from a default set of publicly available water controls. From the selected set of samples, the distribution of reads for each taxon is computed. The z-score field of the IDseq sample report is calculated as the z-score for each taxonomic ID based on its prevalence in the selected background model. Specifically, the z-score for a taxon in sample A is computed as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{(\text{rpm of taxon in sample}) - \text{mean (rpm of taxon in background model samples)}}{\text{std\_dev (rpm of taxon in background model samples)}}.$$

Thus, if a taxon is present at higher abundance in the sample than the controls, it will have a z-score >1. If a particular taxonomic ID is not found in the set of control samples, then the z-score will be set to 100. If the taxonomic ID is not found in the sample, the z-score will be set to -100. The z-score metric also feeds into the "aggregate score," which combines information from NT rPM, NR rPM, NT z-score, and NR z-score to provide an estimate of "microbial importance" for a particular sample

based on the relative abundance both within the sample as well as in the background. This experimental metric aims to rank rare organisms that may be implicated in an infection higher, even if they are present only at low abundance.

External benchmarks—datasets and metrics

Datasets evaluated by Ye et al. [46] in their benchmark analysis of 20 mNGS tools were downloaded from google storage location <gs://metax-bakeoff-2019>. The raw .fastq files were uploaded to IDseq (Supplementary Table S2). The truth files for each of the datasets were obtained from [70] and are available in the Notes field of the IDseq metadata. The code developed by Ye et al. was downloaded from the GitHub repository. IDseq sample reports were downloaded upon completion and processed to produce species-level relative abundance estimates for each sample—specifically, the proportion of total reads (by NT and NR) was computed and used as input to the script. The IDseq results were processed in parallel with the data analyzed for the Ye et al. article. The scripts used to run this analysis are available as well [71]. Modifications to the original script are annotated as "##IDseq EDIT." The computed metrics (AUPR, L2 distance, precision, recall, and f1-score) were then output as .csv files and plotted (Fig. 3, Supplementary Fig. S3).

idseq-bench: IDseq benchmarking tool

The idseq-bench tool [72] was developed as a resource to enable the IDseq team to benchmark datasets internally [73]. The tool is open source and available for external users to generate benchmarks appropriate for their particular use case. Full documentation can be found on GitHub. Briefly, the tool enables users to simulate NGS sequencing data from known microbes. By indicating the GenBank reference accession, idseq-bench uses the InSilicoSeq simulation tool [74] to generate reads in accordance with known sequencing error models. The true organism from which each read was simulated contains a tag indicating the known accession and species-, genus-, and family-level taxonomic IDs. The idseq-bench tool then uses this information to characterize performance of the IDseq pipeline results. The tool provides metrics for read-level recall at the species, genus, and family level, as well as sample-level AUPR, L2, precision, recall, etc. For samples that were not simulated internally, the tool enables users to supply a gold standard file (comparable to those obtained for the Cell Benchmarks Datasets) and compute sample-level metrics against that file.

Internal benchmarks—divergent virus simulation and analysis

A reference genome for Rhinovirus C (RefSeq NC_009996.1) was identified and the associated coding sequence .fasta file was downloaded from RefSeq (RefSeq, RRID:SCR_003496). VIRAPOPS forward viral simulation [75] was used to simulate 5,000 generations of viral evolution using default parameters. From the simulated data, sequences were selected at intervals of 5% nucleotide sequence identity to the original reference and compiled into a fasta file. This was then used as input to the idseq-bench simulation tool for benchmark simulation, which used InSilicoSeq [74] to simulate 10,000 sequencing reads of length 126 for each divergent virus genome according to a HiSeq error model. This resulted in 195.8× coverage of each divergent viral genome, consistent with the relatively high coverage of viral genomes seen by IDseq analysis of samples with high viral load. The simulated

fastq files were then uploaded to IDseq project HRhinoC Simulation (Samples HRC_100, HRC_99, HRC_95, … HRC_025, Supplementary Table S3).

To evaluate the limit of detection for divergent viruses, the total recall of Rhinovirus C reads was evaluated at each level of simulated divergence, for each tool. Additionally, the number of reads aligning to false-positive species was tracked. Offline analysis was done using the contigs generated by IDseq for samples where IDseq failed to identify Rhinovirus C. For simulated samples HRC_070 through HRC_025, the "unmapped contigs" were downloaded and aligned via BLASTx in the NCBI BLAST web interface using default parameters [40]. Samples for which the BLASTx result returned Rhinovirus C were marked as "potentially possible" and the greatest level of divergence was recorded.

Internal benchmarks—running Kraken2

To compare internal benchmark samples against Kraken2 (Kraken, RRID:SCR_005484) [15], a Kraken2 database was generated from the NCBI NT sequence database [76]. The following command line parameters were used to download and build the reference database. Finally, simulated sequencing files were run via the following commands.

Download the NCBI Database:
kraken2-build –download-library nt –db db_ncbi_nt
Build the Kraken2 NCBI Database:
kraken2-build –build –db db_ncbi_nt –threads 8
Run Kraken2 on benchmark datasets:
# classify: running kraken changes slightly based on the sample being compressed/decompressed or single/double pair
BENCHMARK = ⟨benchmark_name_minus-R#⟩ FORMAT = fastq bash -c '/usr/local/sbin/kraken2 –db databases/kraken2/ncbi_nt –threads 8 –gzip-compressed –classified-out results/kraken2/$BENCHMARK.classified_seqs#.fq –unclassified-out results/kraken2/$BENCHMARK.unclassified_seqs#.fq –output results/kraken2/$BENCHMARK.kraken2.out –paired benchmarks/${BENCHMARK}_R1.$FORMAT.gz benchmarks/${BENCHMARK}_R2.$FORMAT.gz &>

Application I—Data processing

In collaboration with Saha et al. [1], 3 samples were identified (CHRF_0000, CHRF_0094, CHRF_0002, from the original NCBI SRA dataset under BioProject PRJNA516582) and re-run on pipeline version 3.13. The pipeline results were filtered using a conservative set of filters, which required NT_rPM > 10 and NT_zscore > 1. The z-score was computed with respect to the public background model CHRF_RNA_Negative, which was used in Saha et al. The background model was generated on the basis of RNA-sequencing data from water samples and negative controls. Metrics were compiled into Table 1 and a heat map was generated using IDseq, with the same filters (Fig. 2D).

Application II—Data processing

In collaboration with Manning et al. [63], RNA was extracted from a sample obtained from a symptomatic patient meeting criteria for possible COVID-19 pneumonia. Libraries were prepared for sequencing as described in Manning et al. and sequenced on an Illumina iSeq100. The raw .fastq files were uploaded to IDseq from the CNM-NIAID laboratory in Phnom Penh, via Illumina BaseSpace, on 31 January 2020 using an NCBI index from September 2019. An NCBI database update was then performed on 2 February 2020 by the IDseq team and the re-

sults were evaluated. These samples were run on IDseq pipeline version 3.18. The data were deposited in public repositories by the original authors and are available at GISAID accession EPI_ISL_411_902. IDseq results for the associated samples are available at [77].

Additional Files

Supplementary Figure S1. The IDseq pipeline visualization.
Supplementary Figure S2. Performance metrics evaluated across 20 mNGS taxonomic identification tools.
Supplementary Figure S3. Per-species recall values for two internal benchmark datsets.
Supplementary Table S1. GitHub repositories contining open-source code for the IDseq pipeline, web application, and benchmarking resources.
Supplementary Table S2. External benchmark datasets and their corresponding IDseq links.
Supplementary Table S3. Internal benchmark datasets and their corresponding IDseq links.
Supplementary Text. Supplemental methods and results associated with 2 benchmark datasets listed in the main text.

Availability of Supporting Source Code and Requirements

Project name: IDseq Portal
Project home page: https://idseq.net
Operating system(s): Platform independent
Programming languages: Python, Ruby, JavaScript
Other requirements: Web browser
License: MIT License
RRID:SCR_019038

Availability of Supporting Data and Materials

Data referenced in this manuscript have been previously published. SRA accession IDs are included in the original publications [1, 46, 63]. Snapshots of the code and tabular data files are available in the GigaDB repository [78].

Abbreviations

ASG: auto-scaling groups; AUPR: area under the precision recall curve; AWS: Amazon Web Services; BLAST: Basic Local Alignment Search Tool; bp: base pairs; CLI: command line interface; CSF: cerebrospinal fluid; DAG: directed acyclic graph; DCR: duplicate compression ratio; EC2: Elastic Compute Cloud; ERCCs: Externa RNA Controls Consortium; kb: kilobase pairs; LZW: Lempel-Ziv-Welch; Mb: megabase pairs; mNGS: metagenomic next-generation sequencing; NCBI: National Center for Biotechnology Information; NIH: National Institutes of Health; NT: Nucleotide, NCBI nucleotide (nt) database; NR: Protein, NCBI non-redundant protein (nr) database; QC: quality control; rPM: reads per million reads sequenced; rRNA: ribosomal RNA; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; SRA: Sequence Read Archive; taxID: taxonomic identifier.

Authors' Contributions

K.L.K. conceived the project. K.L.K. and J.L.D. structured the draft and provided final editing. K.L.K. coordinated and drafted the manuscript and synthesized comments provided by all au-

Acknowledgements

References

1. Saha S, Ramesh A, Kalantar K, et al. Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive chikungunya virus outbreak and other unrealized pathogens. MBio 2019;**10**(6):e02877–19.

2. Simner PJ, Miller S, Carroll KC. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. Clin Infect Dis 2018;**66**:778.

3. Lu J, Breitwieser FP, Thielen P, et al. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci 2017;**3**:e104.

4. Kim D, Song L, Breitwieser FP, et al. Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Res 2016;**26**:1721–9.

5. Walker MA, Pedamallu CS, Ojesina AI, et al. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics 2018;**34**(24):4287–9.

6. Břinda K, Salikhov K, Pignotti S, et al. karel-brinda/prophyle: ProPhyle 0.3.1.0. Zenodo 2017, doi:10.5281/zenodo.1054443.

7. Corvelo A, Clarke WE, Robine N, et al. taxMaps: Comprehensive and highly accurate taxonomic classification of short-read data in reasonable time. Genome Res 2018;**28**:751–8.

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

9. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;**7**:11257.

10. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 2016;**32**(9):1323–30.

11. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015;**12**:902–3.

12. Milanese A, Mende DR, Paoli L, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 2019;**10**:1014.

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

14. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics 2016;**32**:3823.

15. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;**20**:257.

16. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;**304**:66–74.

17. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. Genome Biol 2018;**19**:198.

18. Ainsworth D, Sternberg MJE, Come R, et al. k-SLAM: Accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. Nucleic Acids Res 2017;**45**(4):1649–56.

19. Morgulis A, Coulouris G, Raytselis Y, et al. Database indexing for production MegaBLAST searches. Bioinformatics 2008;**24**:1757–64.

20. Liu X, Yu Y, Liu J, et al. A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. Bioinformatics 2018;**34**:171–8.

21. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. Annu Rev Pathol Mech Dis 2019;**14**:319–38.

22. One Codex. https://www.onecodex.com/. Accessed 11 March 2020.

23. Clarke EL, Taylor LJ, Zhao C, et al. Sunbeam: An extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome 2019;**7**:46.

24. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 2014;**24**:1180–92.

25. IDseq Portal. https://idseq.net. Accessed April 2020.

26. Yozwiak NL, Skewes-Cox P, Stenglein MD, et al. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis 2012;**6**:e1485.

27. IDseq Help Center. https://help.idseq.net. Accessed April 2020.

28. Dobin A, Davis CA, Schlesinger F, et al. Sequence analysis STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013;**29**:15–21.

29. Langelier C, Kalantar KL, Moazed F, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A 2018;**115**:E12353–62.

30. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 2014;**30**:2114–20.

31. Ruby JG, Bellare P, DeRisi JL. PRICE: Software for the targeted assembly of components of (meta) genomic sequence data. G3 (Bethesda) 2013;**3**:865–80.

32. Li Ã W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;**22**:1658–9.

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**:357–9.

34. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 2010;**26**:873–81.

35. Davis NM, Proctor DiM, Holmes SP, et al. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018;**6**:226.

36. Zinter MS, Mayday MY, Ryckman KK, et al. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome 2019;**7**:62.

37. NCBI nucleotide database. ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/.Accessed 2020.

38. Ye Y, Choi JH, Tang H. RAPSearch: A fast protein similarity search tool for short reads. BMC Bioinformatics 2011;**12**:159.

39. Index of /pub/taxonomy/accession2taxid. ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid. Accessed 26 February 2020.

40. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol 1990;**215**:403–10.

41. Kulikov AS, Prjibelski AD, Tesler G, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;**19**:455–77.

42. IDseq Heatmap. https://idseq.net/zlfl1. Accessed April 2020.

43. IDseq DAG GitHub Repository. https://github.com/chanzuckerberg/idseq-dag. Accessed April 2020.

44. s3mi GitHub Repository. https://github.com/chanzuckerberg/s3mi. Accessed April 2020.

45. IDseq web GitHub Repository. https://github.com/chanzuckerberg/idseq-web. Accessed April 2020.

46. Ye SH, Siddle KJ, Park DJ, et al. Benchmarking metagenomics tools for taxonomic classification. Cell 2019;**178**: 779–94.

47. Luo C, Knight R, Siljander H, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 2015;**33**:1045–52.

48. Scholz M, Ward DV, Pasolli E, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 2016;**13**:435–8.

49. Truong DT, Tett A, Pasolli E, et al. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 2017;**27**:626–38.

50. Steinhauer DA, Holland JJ. Rapid evolution of RNA viruses. Annu Rev Microbiol 1987;**41**:409–31.

51. Woolhouse MEJ, Brierley L, McCaffery C, et al. Assessing the epidemic potential of RNA and DNA viruses. Emerg Infect Dis 2016;**22**:2037–44.

52. Schuffenecker I, Iteman I, Michault A, et al. Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. PLoS Med 2006;**3**:e263.

53. Pu J, Wang S, Yin Y, et al. Evolution of the H9N2 influenza genotype that facilitated the genesis of the novel H7N9 virus. Proc Natl Acad Sci U S A 2015;**112**:548–53.

54. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;**382**:727.

55. Chappell JG, Byaruhanga T, Tsoleridis T, et al. Identification of infectious agents in high-throughput sequencing data sets is easily achievable using free, cloud-based bioinformatics platforms. J Clin Microbiol 2019;**57**:e01386–19.

56. Ramesh A, Nakielny S, Hsu J, et al. Metagenomic next-generation sequencing of samples from pediatric febrile illness in Tororo, Uganda. PLoS One 2019;**14**: e0218318.

57. Crawford E, Kamm J, Miller S, et al. Investigating transfusion-related sepsis using culture-independent metagenomic sequencing. Clin Infect Dis 2020;**71**(5):1179–85.

58. Hasan MR, Sundararaju S, Tang P, et al. A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting. Sci Rep 2020;**10**(1): 11194.

59. Ruppé E, Schrenzel J. Messages from the second International Conference on Clinical Metagenomics (ICCMg2). Microbes Infect 2018;**20**:222–7.

60. Wilson MR, O'Donovan BD, Gelfand JM, et al. Chronic meningitis investigated via metagenomic next-generation sequencing. JAMA Neurol 2018;**75**:947–55.

61. Mollerup S, Friis-Nielsen J, Vinner L, et al. *Propionibacterium acnes*: Disease-causing agent or common contaminant? detection in diverse patient samples by next-generation sequencing. J Clin Microbiol 2016;**54**:980–7.

62. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One 2014;**9**:e97876.

63. Manning JE, Bohl JA, Lay S, et al. Rapid metagenomic characterization of a case of imported COVID-19 in Cambodia. bioRxiv 2020, doi:10.1101/2020.03.02.968818.

64. Retallack H, Okihiro MS, Britton E, et al. Metagenomic next-generation sequencing reveals *Miamiensis avidus* (Ciliophora: Scuticocilitida) in the 2017 epizootic of leopard sharks (*Triakis semifasciata*) in San Francisco Bay, California, USA. J Wildl Dis 2019;**55**:375.

65. Batson J, Dudas G, Haas-Stapleton E, et al. Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter. bioRxiv 2020, doi:10.1101/2020.02.10.942854.

66. Sczyrba A, Hofmann P, Belmann P, et al. Critical assessment of metagenome interpretation - A benchmark of metagenomics software. Nat Methods 2017;**14**:1063–71.

67. McIntyre ABR, Ounit R, Afshinnekoo E, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 2017;**18**:182.

68. IDseq Workflows Wiki.https://github.com/chanzuckerberg/idseq-workflows/wiki. Accessed August 2020.

69. Zinter MS, Dvorak CC, Mayday MY, et al. Clinical infectious diseases pulmonary metagenomic sequencing suggests missed infections in immunocompromised children. Clin Infect Dis 2019;**68**:1847.

70. Metax Bakeoff GitHub Repository. https://github.com/yesimon/metax_bakeoff_2019. Accessed January 2020.

71. IDseq benchmark manuscript GitHub Repository. https://github.com/katrinakalantar/idseq-benchmark-manuscript. Accessed April 2020.

72. IDseq bench GitHub Repository. https://github.com/chanzuckerberg/idseq-bench. Accessed April 2020.

73. GitHub - chanzuckerberg/idseq-bench: IDseq infectious disease benchmarking tools. https://github.com/chanzuckerberg/idseq-bench. Accessed 11 March 2020.

74. Gourlé H, Karlsson-Lindsjö O, Hayer J, et al. Simulating Illumina metagenomic data with InSilicoSeq. https://github.com/HadrienG/InSilicoSeq. Accessed 26 February 2020.

75. Petitjean M, Vanet A. VIRAPOPS2 supports the influenza virus reassortments. Biol Med 2014;**9**:18.

76. O'leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2015;**44**:733–45.

77. COVID-19 Case in Cambodia. http://public.idseq.net. Accessed April 2020.

78. Kalantar KL, Carvalho T, deBourcy CFA, et al. Supporting data for "IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring." GigaScience Database 2020, http://dx.doi.org/10.5524/100803.