(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 January 2025 (02.01.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/007145 A2

(51) International Patent Classification:
C12Q 1/689 (2018.01) *G16B 30/00* (2019.01)

(21) International Application Number:
PCT/US2024/036435

(22) International Filing Date:
01 July 2024 (01.07.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/511,532 30 June 2023 (30.06.2023) US

(71) Applicant: **THE REGENTS OF THE UNIVERSITY OF CALIFORNIA** [US/US]; 1111 Franklin Street, Twelfth Floor, Oakland, CA 94607-5200 (US).

(72) Inventors: **ESKIN, Eleazar**; 3011 2nd Street, Santa Monica, CA 90405 (US). **ARBOLEDA, Valerie**; 1111 Franklin Street, Twelfth Floor, Oakland, CA 94607-5200 (US).

(74) Agent: **DENG, Yingxin**; KPPB LLP, 3780 Kilroy Airport Way, Suite 320, Long Beach, CA 90806 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: METAGENOMIC DIAGNOSTIC SYSTEMS AND METHODS THEREOF

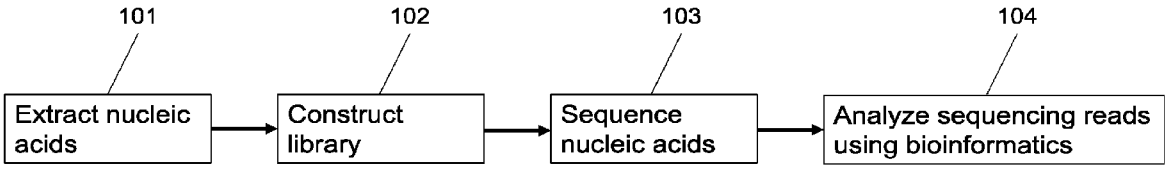


FIG. 1

(57) Abstract: Systems and methods for SwabSeq metagenomic diagnostic platforms are described. The metagenomic diagnostic platforms are meta-genomics based untargeted next generation sequencing (NGS) diagnostic systems for detecting nucleic acids from pathogens. Samples undergo nucleic acid extraction followed by NGS library preparation and sequencing. Results of sequencing can be processed by bioinformatics pipelines. Automation is used to increase accuracy and analyze up to thousands of samples simultaneously without contamination.

WO 2025/007145 A2

Metagenomic Diagnostic Systems and Methods Thereof

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The current application claims the benefit of and priority under 35 U.S.C. § 119 (e) to U.S. Provisional Patent Application No. 63/511,532 entitled "Metagenomic Diagnostic Systems and Methods Thereof" filed June 30, 2023. The disclosure of U.S. Provisional Patent Application No. 63/511,532 is hereby incorporated by reference in its entirety for all purposes.

STATEMENT OF FEDERAL FUNDING

[0002] This invention was made with government support under 75A50122C00033 awarded by the U.S. Department of Health and Human Services. The government has certain rights in the invention.

SEQUENCE LISTING

[0003] The instant application contains a Sequence Listing which has been submitted electronically in XML format and is hereby incorporated by reference in its entirety. Said XML copy, created on July 1, 2024, is named R4-08645.PCT.xml and is 3 kilobytes in size.

FIELD OF THE INVENTION

[0004] The present invention generally relates to systems and methods for metagenomic diagnosis.

BACKGROUND

[0005] Culturing is a conventional method to identify pathogens. However, culturing is time-consuming and many pathogens that need specific culture conditions are difficult to grow. Metagenomic sequencing is a target-independent approach that offers a comprehensive view of the pathogenic agents and provides a detection of common and unexpected pathogens in samples. This technology performs well in detecting rare pathogens and/or novel viral strains. Current approaches to metagenomic sequencing

can only process a small number of samples at a time and each sample costs hundreds of dollars. There are multiple challenges to processing larger amounts of samples including difficulty to prepare the samples and cross-contamination which is when the samples are mixed during processing causing incorrect results.

BRIEF SUMMARY

[0006] Many embodiments are directed to systems and methods for metagenomic diagnosis.

[0007] Some embodiments include a method of detecting a pathogen comprising, obtaining a plurality of nucleic acids from a sample; sequencing the plurality of nucleic acids to obtain a plurality of reads; analyzing the plurality of reads using a bioinformatic process; and identifying the pathogen from the analyzed reads.

[0008] Some embodiments further comprise obtaining the sample using at least one of: a swab, a Q-tip, a wipe, and a cloth.

[0009] In some embodiments, the sample is a tissue, an organ, a bodily fluid, an object, a surface, or a container.

[0010] In some embodiments, the sample is at least one of: a turbinate swab, a nasal pharyngeal swab, and saliva.

[0011] In some embodiments, the sample is a clinical sample; wherein the clinical sample is at least one of: urine, a tissue, a skin swab, bronchoalveolar lavage (BAL), sputum, and blood.

[0012] In some embodiments, the sample is an upper respiratory specimen or a lower respiratory specimen from a subject.

[0013] In some embodiments, the obtaining step obtains the plurality of nucleic acids using at least one nucleic acid extraction kit.

[0014] Some embodiments further comprise adding a known quantity of Bacteriophage MS2 to the sample.

[0015] Some embodiments further comprise constructing at least one next generation sequencing library.

[0016] In some embodiments, the at least one next generation sequencing library is at least one of: a DNA library, and an RNA library.

[0017] In some embodiments, the sequencing step uses the at least one next generation sequencing library.

[0018] In some embodiments, the analyzing step comprises matching a plurality of sequencing reads to at least one pathogen sequence in a database.

[0019] Some embodiments further comprise generating a plurality of simulated reads using a set of curated pathogen sequences; analyzing the plurality of simulated reads using the database to check if the database contains an error; and eliminating a plurality of false reads in the database by analyzing the plurality of simulated reads.

[0020] Some embodiments further comprise using the plurality of reads of the pathogen and a plurality of reads of a control added prior to nucleic acid extraction or a plurality of reads of a control added after nucleic acid extraction to quantify an amount of the sample.

[0021] In some embodiments, the control is Bacteriophage MS2.

[0022] In some embodiments, the pathogen is a pathogenic organism, a bacterium, a virus, or a fungus.

[0023] Some embodiments further comprise processing a plurality of samples at a time using automation.

[0024] Some embodiments further comprise processing a plurality of samples at a time using automation by interleaving a plurality sets of samples to increase a number that is processed on an automated instrument.

[0025] Some embodiments further comprise processing a plurality of samples using at least one microfluidic liquid handler.

[0026] In some embodiments, automation eliminates cross-sample contamination when processing the plurality of samples.

[0027] Some embodiments further comprise adding at least one molecular barcode to each of the plurality of samples, and mixing the plurality of samples.

[0028] In some embodiments, more than 90 samples are processed together at a time.

[0029] In some embodiments, more than 200 samples are processed together at a time.

[0030] In some embodiments, more than 1000 samples are processed together at a time.

[0031] In some embodiments, the preparation for sequencing and analyzing is performed on a microfluidic liquid handler.

[0032] Some embodiments further comprise a depletion process configured to improve a performance of the mixture of barcoded plurality of samples.

[0033] Some embodiments further comprise combining a plurality of targeted diagnostic samples with a plurality of metagenomic diagnostic samples in a same sequencing process.

[0034] Some embodiments further comprise using a nucleic extraction process that is optimized for extracting bacteria and viruses compared to host nucleic acids.

[0035] In some embodiments, sample to result using short read sequencers is completed in less than 11 hours.

[0036] Some embodiments include a method of a bioinformatic analysis comprising, generating a plurality of simulated genomic reads using a plurality of pathogen genomic sequences; eliminating a plurality of false reads in a database by analyzing the plurality of simulated genomic reads; comparing a plurality of sequencing reads from a sample with the database; and identifying a pathogen from the sample; wherein the eliminated false reads in the database improves identification accuracy.

[0037] Some embodiments further comprise analyzing the plurality of simulated genomic reads using the database to check if the database contains an error.

[0038] Some embodiments further comprise generating a plurality reads of a known quantity of Bacteriophage MS2 and using the plurality reads of the known quantity of Bacteriophage MS2 and the plurality of sequencing reads to quantify an amount of the sample.

[0039] Some embodiments further comprise extracting a plurality of nucleic acids from the sample.

[0040] In some embodiments, the extracting step uses at least one nucleic acid extraction kit.

[0041] Some embodiments further comprise using a next generation sequencer to generate the plurality of sequencing reads.

[0042] In some embodiments, the pathogen is a bacterium or a virus.

[0043] In some embodiments, the pathogen is reconstructed using a genome assembly technique.

[0044] In some embodiments, eliminating the plurality of false reads uses a manual process.

[0045] In some embodiments, eliminating the plurality of false reads uses a computer assisted process.

[0046] In some embodiments, the computer assisted process uses artificial intelligence.

[0047] Some embodiments further comprise using a control sequence that is an organism other than Bacteriophage MS2.

[0048] Some embodiments further comprise adding a control RNA or DNA after nucleic acid extraction to generate a plurality of simulated genomic reads.

[0049] In some embodiments, the database used has publicly available sequencing data that is curated prior to analysis using efficient algorithms.

[0050] Additional embodiments and features are set forth in part in the description that follows, and in part will become apparent to those skilled in the art upon examination of the specification or may be learned by the practice of the disclosure. A further understanding of the nature and advantages of the present disclosure may be realized by reference to the remaining portions of the specification and the drawings, which forms a part of this disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0051] The description will be more fully understood with reference to the following figures, which are presented as exemplary embodiments of the invention and should not be construed as a complete recitation of the scope of the invention. It should be noted that the patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0052] Figure 1 illustrates a process of SwabSeq metagenomics diagnostic platform workflow in accordance with an embodiment.

[0053] Figure 2 illustrates a 96 well plate with the control positions marked in accordance with an embodiment.

[0054] Figure 3 illustrates workflows for the metagenomic diagnostic platform multiplex assay in accordance with an embodiment.

[0055] Figure 4 illustrates a process of bioinformatic analysis process in accordance with an embodiment.

[0056] Figure 5 illustrates sample output for 20 COVID-19 samples analyzed using analyzed using the metagenomic diagnostic platform in accordance with an embodiment.

[0057] Figure 6A shows True Positive positions of Parainfluenza 3 in accordance with an embodiment.

[0058] Figure 6B shows predicted positions using manual library preparation protocol in accordance with an embodiment.

[0059] Figure 6C shows predicted predictions using automated library preparation protocol in accordance with an embodiment.

[0060] Figures 7A and 7B illustrate interleaving of two plates in accordance with an embodiment.

[0061] Figure 8A through 8I show the results from automation validation for three additional pathogens in accordance with an embodiment.

DETAILED DESCRIPTION

[0062] Identifying different pathogens in a sample may be challenging. These challenges stem from the fact that the number of sequencing reads that correspond to a pathogen is a very small fraction of the total amount of sequence obtained for a sample, often in the range of 1 in about 10,000 reads and in advance the pathogens present in the sample are unknown species.

[0063] In order to identify a pathogen, sequencing reads corresponding to a pathogen genome need to be identified, and sequencing reads that do not correspond to a pathogen genome also need to be identified. One approach would be to obtain a set of pathogen genomes and compare each sequencing read to each of the genomes. If there is a close match, the read can be assigned to the pathogen. Once this is completed, the counts of reads assigned to pathogen genomes can be used as an indicator of which

genomes are present in the sample. However, this approach can result in false positives. Many pathogens share some genetic sequence. Other genomes such as human genomes or other organisms may be present in the sample. This can result in false positives where a read is assigned to a pathogen even though it originates from a different genome. This is especially the case when an organism that is related to the pathogen is present in the sample.

[0064] Systems and methods of metagenomic diagnostic platforms are described. Metagenomic agnostic diagnostic platform is a next generation sequencing (NSG) based diagnostic that can detect nucleic acids from various pathogens, and identify various pathogens present in samples. Examples of pathogens can include (but are not limited to) bacteria and/or viruses. Pathogens can be detected in upper respiratory specimens from patients suspected of infection. Assays can be optimized to improve pathogen identification in clinical samples. To prepare for identification, samples undergo nucleic acid extraction followed by NGS library preparation and sequencing. An internal control which is the Bacteriophage MS2 is added to each sample prior to extraction. Results of sequencing can be processed by bioinformatics analysis (or pipelines) which can perform (but not limited to) quality controls, checks for internal controls, and match reads to pathogens. In many embodiments, workflows to return results can be achieved within 24 hours. In several embodiments, the metagenomic diagnostic platforms can analyze 5 samples in less than about 12 hours; or 96 or 192 samples in about 24 hours. Some embodiments use a MiniSeq Rapid Run Kit that can analyze 5 samples in less than about 12 hours. Some embodiments use an Illumina® NextSeq 2000 Sequencer that can analyze about 96 to 192 samples in about 24 hours at a cost of under $50 per sample. When applied to even larger numbers of samples simultaneously analyzed, this price can approach about $10 per sample. Many embodiments provide that existing approaches cannot process even close to that many samples and are much more expensive for each sample.

Table 1. Turnaround times and sequencing costs per sample for configurations of the SwabSeq Metagenomic Diagnostic Platform. With planned modifications to the

automated workflow, the sample to answer time will reduce to 23.1 hours for 96 samples and 24.6 hours for 192 samples

	MiniSeq 5	NextSeq 2000 96	NextSeq 2000 192
Sequencer	MiniSeq	NextSeq 2000	NextSeq 2000
Flow cell	Rapid 50bp	P2 2 x 50bp	P3 1 x 50bp
Automated/Manual	Manual	Automated	Automated
Extraction Time	1 hour	1 hour	1 hour
Library Prep Time	6 hours	9.8 hours	14.5 hours
Sequencing Time	3.3 hours	13 hours	11 hours
Bioinformatics Time	0.2 hours	0.3 hours	0.4 hours
Sample to Result Time	10.5 hours	24,1 hours	26.9 hours
Sequencing $ / Sample	$200	$15	$15

[0065] Metagenomic diagnostic platforms (also referred to as SwabSeq) do not perform targeted amplification of pathogen sequence such as (but not limited to) PCR. Instead, the platforms obtain hundreds of thousands of sequencing reads per sample, most of them corresponding to patients and/or host nucleic acids. The platforms then use a bioinformatics pipeline to identify and classify the small fraction of the reads that correspond to pathogen sequence using nuclei acid databases such as (but not limited to) annotated databases, Refseq database, and/or GenBank database. Prior to classification, a cross-reactivity analysis can be performed using curated genome sequences to verify that the pipeline can accurately identify the target pathogens.

[0066] Figure 1 illustrates a process of SwabSeq metagenomics diagnostic platform workflow in accordance with an embodiment. The process starts with extracting (101) nucleic acid from a sample. Samples can be collected using a swab, a Q-tip, a wipe, and/or a cloth, from a source such as (but not limited to) an organ, a tissue, a liquid, a bodily fluid, an organism, a living species, an animal, a human, a patient, an object, a surface, a container, and/or any of a source that may contain pathogens. Some embodiments use turbinate swabs; or nasal pharyngeal swabs; or saliva. The collected sample can be prepared (if needed) or used directly for nucleic acid extraction. Nucleic acid extraction can be performed using various extraction kits. Nucleic acid extraction for various specimen types can be performed using instruments such as (but not limited to) Thermo Fisher® KingFisher Apex Instrument (5400930) and Zymo Quick-DNA/RNA Viral

Magbead (R2141). Sequencing libraries can be created with the extracted RNA using the NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® (E6420L).

[0067] In some embodiments, prior to extraction, a control which is fixed quantity of Bacteriophage MS2 is added to each sample. This control is later used to quantify the amount of pathogen in the sample. In some embodiments, an additional control of RNA or DNA is added to the extracted RNA or RNA from the sample and used as control and to further quantify the amount of pathogen in the sample as well as potentially identify issues of contamination.

[0068] Construct (102) library with extracted nucleic acids. Constructing RNA and/or DNA library for NGS analysis can include the steps: fragmenting and/or sizing the target sequences to a desired length; converting target to double-stranded DNA; attaching oligonucleotide adapters to the ends of target fragments; and quantitating the final library product for sequencing. The library preparation can be automated or manual.

[0069] Sequence (103) nucleic acids using the prepared library. Sequencing reads and/or data can be generated using various sequencers such as (but not limited to) Illumina® sequencers, Oxford Nanopore® sequencers, and/or Qiagen® sequencers. Some embodiments use an Illumina® Sequencer (NextSeq 2000) in Illumina® BCL format.

[0070] Analyze (104) sequencing reads using metagenomic diagnostic platform bioinformatics pipeline. Analysis results output identification of a single pathogen and/or a plurality of pathogens that the extracted nucleic acids are from. Bioinformatic analysis of metagenomics diagnostic platforms is discussed in further details below.

[0071] In many embodiments, samples are organized into a 96 well plate where at least 2 of the positions in the plate correspond to controls. One control is a no-template control containing nuclease free water. One control is a positive control which is a contrived sample containing a mixture of target pathogens. Figure 2 illustrates a 96 well plate with the control positions marked in accordance with an embodiment. The control positions are always one of four positions on the 96 well plate (A01, A02, H07, H08). Prior to RNA extraction, about 10^6 PFU of Bacteriophage MS2 is added to each sample and control as an additional positive control. Some embodiments add a control after nucleic acid extraction.

[0072] Figure 3 illustrates workflows for the metagenomic diagnostic platform multiplex assay in accordance with an embodiment. Collecting (301) samples such as extracting nucleic acid from a sample. Samples can be swabbed in about 3 mL viral transport medium (VTM) or universal transport medium (UTM). Extract (302) RNA from the samples. The library can be prepared manually (303) or automatically (304). Automated library preparation can include (but not limited to) automated Biomek i7. Sequence (305) the extracted samples using various sequencing techniques or platforms. Analyze (306) the sequencing reads using (but not limited to) RefSeq bioinformatics database.

[0073] Many embodiments identify pathogens using a database of all available genome sequences and assigning each read to the genome that is its closest match. Using a database with all available genome sequences instead of a set of pathogen genomes can enhance accuracy in pathogen identification. Several embodiments keep track of reads that match multiple genomes and consider the set of genomes that match for further analysis. Such computational analysis of sequencing data for pathogen identification can improve accuracy as the information in the other genomes can accurately identify if a read comes from a different organism. Prior methods for pathogen identification may break up a genome into pieces and match each piece to an indexed database of genomes. There can be two drawbacks with prior methods. The first is that databases themselves may have errors which can lead to inaccurate predictions. The second is that some reads or pieces of reads match multiple genomes which creates computational difficulties for building the indexed database of genomes and performing the matching.

[0074] To address database errors, metagenomic diagnostic platforms in accordance with many embodiments perform a preprocessing step that utilizes a set of curated pathogen sequences. These sequences can be used to generate simulated reads which are then analyzed using computational analysis in accordance with several embodiments. Some embodiments then identify any reads and/or pieces of reads that are incorrectly classified using the database. Certain embodiments can then mark these regions and will ignore them if they show up in an actual sample. Same analysis approaches can be applied using genomes of organisms that are likely to be present in clinical samples. After performing simulated read analysis, regions of the genome that are ambiguous and/or

can lead to errors can be eliminated, such that the accuracy of the computational analysis improves.

[0075] To address the issue of pieces of reads matching multiple genomes, many embodiments allow database index to match multiple genomes for each piece of the genome considered. This may need solving a computational challenge of building an index for the genomes that allows for this. Several embodiments develop algorithms for constructing such an index which focuses on the pathogen and related genomes. Such algorithms can keep track of all the genomes that a piece of a sequence matches. In some embodiments, the set of genomes can be broken up into subsets. Certain embodiments use k-mer analysis algorithms that can compute an exact matrix of which piece matches which genome for a subset of the genomes and then combine these matrices. The algorithm in accordance with some embodiments may not compute the complete matrix because the complete matrix is too large to store in memory, but it is possible to compute an approximation that will provide high accuracy in terms of classification.

[0076] To further increase accuracy, several embodiments take all the classified reads and their potential genomes and align the reads to the gnomes. By examining the quality of the alignments, ambiguities may be solved to further increase accuracy.

[0077] In several embodiments, sequencing data can be analyzed using the metagenomic diagnostic platforms that have 3 phases: fastq generation and quality control; read classification into genome of origin; internal controls and diagnostic calling. Figure 4 illustrates a process of bioinformatic analysis process in accordance with an embodiment of the invention. The SwabSeq metagenomic diagnostic platform bioinformatics analysis process can include the following steps. Convert BCL format data to fastq format data using Illumina® Dragen software. Apply (401) quality control by filtering the fastq files using fastp. The BCL data can be converted to paired fastq format using the bcl-convert software from Illumina® and a sample sheet created with barcode information obtained from New England Biolab®. Quality control can be performed using the fastp software using default settings except for passing the -q 7 option which reduces the sequencing quality level slightly compared to the default setting. This optimization can

be determined by experiments over contrived and clinical samples. Reads can be filtered out if:

1) 40% or more of the positions in the read have a PHRED quality score of less than 7;
2) the read has a complexity of less than 30%;
3) the read has less than 15 positions remaining after removing any adapter sequence.

The result of this process is a pair of qc filtered fastq files for each sample.

[0078] Classify remaining reads into taxa using kraken 2 and match (402) reads to databases such as RefSeq database. The database can include RefSeq archaea, bacteria, viral, plasmid, human and UniVec_Core from National Center for Biotechnology Information (NCBI). Any pathogen target genomes which are not present in the standard database can be added to the database. Previously, curated pathogen genomes (such as from the ARGOS database) were utilized to validate the pipeline and identify any regions of the RefSeq genome that caused misclassified reads. Any reads that mapped to these regions are removed. Reads that map to multiple target pathogens are removed from the analysis. Line data consisting of counts of unclassified reads and classified reads for human, bacteria, viruses and counts of reads that uniquely map to each respiratory pathogen target are extracted from the kraken 2 results. This line data is evaluated using internal control checks and used to generate diagnostic calls.

[0079] The software kraken 2 is run using default parameters. A set of read counts for a specific set of genomes with taxonomic identifiers in Table 2 is extracted from the results.

Table 2 lists taxonomic Identifiers with read counts extracted in SwabSeq Metagenomic Bioinformatics Pipeline. The taxonomies that correspond to pathogen targets are marked in the Pathogen Target column. Some of the pathogen sequences are labeled with a "Pathogen Group" which are closely related sequences that are difficult to distinguish. For these sequences, the diagnostic calling algorithm reports the presence or absence of a sequence from the group (e.g., "Influenza A") since there is not often enough information to identify the actual sequence. Pathogen group entries with an asterisk denote sequences where contrived sequences were available to validate the detection of those specific genomes which allow the possibility of the diagnostic algorithm predicting the specific sequence (e.g., Influenza A H1N1).

Taxonomic ID	Genome Name	Pathogen Target	Pathogen Group
0	unclassified		
1	classified		
9606	human		
2	bacteria		
10239	virus		
12022	Bacteriophage MS2	Control	
290028	Human coronavirus HKU1	P	
31631	Human coronavirus OC43	P	
277944	Human coronavirus NL63	P	
11137	Human coronavirus 229E	P	
162145	Human metapneumovirus	P	
12059	Enterovirus	P	Enterovirus
42789	Enterovirus D68	P	Enterovirus
147711	Rhinovirus A	P	Enterovirus
147712	Rhinovirus B	P	Enterovirus
463676	Rhinovirus C	P	Enterovirus
11250	Human orthopneumovirus	P	RSV
12814	RSV	P	RSV
11320	Influenza A	P	Influenza A
119210	Influenza A H3N2	P	Influenza A*
114727	Influenza A H1N1	P	Influenza A*
11520	Influenza B	P	
2697049	SARS-CoV-2	P	
12730	Parainfluenza 1	P	
2560525	Parainfluenza 2	P	
11216	Parainfluenza 3	P	
11224	Parainfluenza 4	P	
10509	Adenovirus	P	
11234	Measles morbillivirus	P	
2560602	Mumps orthorubulavirus	P	

[0080] The results of the previous steps are a count of the number of reads for each taxonomic identifier in Table 1. "classified" is the sum of reads that correspond to any taxonomic identifier and "unclassified" are reads which do not correspond to any taxonomy identifier. Apply (403) three internal control checks to kraken 2 results to verify that a sample is valid:

1) Sequence Quality (Internal Control 1): Checks the sequence quality by verifying that the percentage of classified reads is greater than 0.5. Passed if the number of classified reads divided by sum of classified and unclassified reads is greater than 0.5.
2) Sequence Depth (Internal Control 2): Checks the sequence quality by verifying that the number of classified reads is over 100,000. Passed if the number of classified reads is over 100,000.
3) Human Reads (Internal Control 3): Checks by verifying if the number of human reads is greater than 20,000. Passed if the number of human reads is greater than 20,000.
4) MS2 Reads (Internal Control 4): Checks by verifying if the number of MS2 reads is greater than 100. Passed is the number of MS2 reads is greater than 100.
5) MS2 Reads Ratio to Total Reads: Checks by verifying if the ratio of MS2 reads is greater than 0.01. Passed if the number of MS2 reads divided by the number of classified reads is greater than 0.01.

[0081] Identify (404) pathogens from reads. A sample that passes all internal control checks is a valid sample. A sample that fails any one of internal control checks is an inconclusive sample. For any valid sample, a pathogen target is considered present in the sample if the number of reads for the pathogen divided by the number of MS2 reads is greater than 0.01. For example, if the number of MS2 reads is 1000, the pathogen read threshold is 10.

[0082] For any pathogen target below that level, the pathogen is considered not present. For pathogen targets not part of "Pathogen Groups" (Table 2), the pathogen results as "detected" if the pathogen is considered present and "not detected" otherwise. For pathogen targets part of pathogen groups, due to sequence similarity between targets within each group, the specific pathogen present is difficult to distinguish. For these targets, if any pathogen from a group is above the threshold, the pathogen group results

as "detected" and otherwise the pathogen group results "not detected". In practice, typically multiple sequences from the pathogen group are above the threshold.

[0083] For pathogen targets listed with an asterisk in their pathogen group (e.g., Influenza A H1N1), contrived samples were available within the pathogen group to enable identification of sequences within the group. For these sequences, the specific pathogen results "detected" if 95% of the reads within the group are from the specific genome.

[0084] In addition, two "plate level" controls must be passed:

1) (NTC Control) Passed if the MS2 reads divided by the total classified reads divided by the total number of classified reads if greater than 0.1 for all of the NTC positions.
2) (Positive Controls) Passed if the pathogen target is "detected" for all pathogens present in the pathogen mixture. The positive control can either originate from a known positive sample or a contrived sample which has a mixture of pathogens.

If either of the "plate level" controls are not passed, the entire plate results as inconclusive.

EXEMPLARY EMBODIMENTS

[0085] Although specific embodiments of systems and apparatuses are discussed in the following sections, it will be understood that these embodiments are provided as exemplary and are not intended to be limiting.

Example 1: SwabSeq Metagenomic Diagnostic Platforms

[0086] The SwabSeq Metagenomic Diagnostic Platform is a meta-genomics based untargeted NGS diagnostic designed to detect nucleic acid from a pathogen in upper respiratory specimens from patients suspected of infection. Samples undergo nucleic acid extraction followed by NGS library preparation and sequencing. Results of sequencing are then processed by a bioinformatics pipeline which performs quality control, checks for internal controls, and matches reads to pathogen. The SwabSeq Metagenomic Diagnostic Platform does not perform any targeted PCR amplification of pathogen sequence. Instead, the platform obtains hundreds of thousands of sequencing reads per sample, most of them corresponding to patient/host nucleic acid. The platform then uses a bioinformatics pipeline to identify and classify the small fraction of the reads that correspond to pathogen sequence using the RefSeq database. Prior to this classification,

a cross-reactivity analysis is performed using curated genome sequences to verify that the pipeline is capable of accurately identifying the target pathogens.

[0087] Several embodiments implement SwabSeq metagenomic diagnostic platform on sample types such as (but not limited to) mid turbinate swabs, nasal pharyngeal swabs and saliva. Sequencing data is analyzed using SwabSeq Metagenomic Diagnostic Pipeline which has 3 phases: Fastq generation and quality control, read classification into genome of origin, and internal controls and diagnostic calling.

[0088] Some embodiments implement Fastq generation and quality control processes. The Illumina® NextSeq 2000 generates the sequencing data in raw BCL format. The BCL data is converted to paired fastq format using the bcl-convert software from Illumina® and a sample sheet created with barcode information obtained from NEB®. Quality control can be carried out using the fastp software using default settings except for passing the -q 7 option which reduces the sequencing quality level slightly compared to the default setting. This optimization can be determined by experiments over contrived and clinical samples. Using fastp, reads are filtered out if: 40% of the positions have a read quality of less than 7; the read has a complexity of less than 30%; the read has less than 15 positions after removing any adapter.

[0089] The result of this process is a pair of qc filtered fastq files for each sample. An additional filtering step is performed to identify filtered reads that ambiguously map to both a pathogen target and either a common contaminant or organism that is often present in respiratory samples. This filtering is performed using the bbduk.sh tool from the BBMap software package. bbduk.sh is run against a “contaminant kmer list” which is a set of kmers from portions of the genomes which are shared between targets and contaminants and have been shown to classify incorrectly. bbduk.sh removes any read that contains one of these kmers. The process to identify the kmers which may cause incorrect classification is described below.

[0090] For each pathogen target, a curated genome can be identified. This is typically a RefSeq genome that is marked as “Complete” or from a reliable source. For SARS-CoV-2, certain embodiments use the RefSeq sequence “NC_045512.2 Severe acute

respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome." The genomes for organisms commonly occurring in nasal samples are included in Table 3.

Table 3. List of organisms commonly present in respiratory samples.

Organisms likely present in respiratory samples
Human
Adenovirus (e.g., C1 Ad. 71)
Human coronavirus 229E
Human coronavirus OC43
Human coronavirus HKU1
Human coronavirus NL63
SARS-CoV-1
MERS-coronavirus
Human Metapneumovirus (hMPV)
Parainfluenza virus 1-4
Influenza A & B
Enterovirus (e.g., EV68)
Respiratory syncytial virus
Rhinovirus
Chlamydia pneumoniae
Haemophilus influenzae
Legionella pneumophila
Mycobacterium tuberculosis
Streptococcus pneumoniae
Streptococcus pyogenes
Bordetella pertussis
Mycoplasma pneumoniae
Pneumocystis jirovecii (PJP)
Candida albicans
Pseudomonas aeruginosa
Staphylococcus epidermis
Streptococcus salivarius

[0091] Some embodiments use the wgsim to generate simulated reads from each organism and run them through the pipeline. For pathogen targets, any reads that are either incorrectly classified as another target pathogen or incorrectly classified as either human or bacteria can be identified. For common organisms, any reads which are incorrectly classified as a target pathogen can be identified. Some embodiments look for kmers of length 31 that are common among the incorrectly classified reads and any kmer that occurs at least 3 times are included in a "contaminant kmers list." The intuition is that

these commonly occurring kmers are characteristic of regions of the genomes that are inherently ambiguous and may lead to incorrect classifications. This list can be used to filter any reads which contain one of these kmers.

[0092] Some embodiments implement the processes for read classification into genome of origin. Reads from each sample are classified into a genome of origin using the kraken 2 software using the "standard" database which consists of RefSeq archaea, bacteria, viral, plasmid, human and UniVec_Core from NCBI. Pathogen target genomes which are not present in the standard database can be added. kraken 2 is run using default parameters. A set of read counts for a specific set of genomes with taxonomic identifiers in Table 1 is extracted from the results.

[0093] Some embodiments provide the processes for internal controls and diagnostic calling. The results of the previous steps are a count of the number of reads for each taxonomic identifier in the table above. "classified" is the sum of reads that correspond to any taxonomic identifier and "unclassified" are reads which do not correspond to any taxonomy identifier. There are five controls which must be passed for a sample to be valid as discussed above. Figure 5 shows sample output for 20 COVID-19 samples analyzed using the SwabSeq Metagenomic Diagnostic Platform in accordance with an embodiment. Each line is the output for one sample. The numbers in each column are the number of reads in each category. For example, for sample 1, 75,441 reads are not able to be classified while 1,354,221 reads are classified. Of these reads, a little over 1.3M are human reads, a little over 40k are bacteria reads and 2,466 are viral reads. Of those reads 2,180 are from SARS-CoV-2. For this sample, each of the three internal controls are passed: 94.7% of the reads are classified to pass Internal Control 1; more than 100,000 reads are classified to pass Internal Control 2; more than 20,000 reads are classified as human to pass Internal Control 3. The fraction of SARS-CoV-2 reads is greater than 0.00001 for samples 1, 2, 11 and 20 which is concordant with PCR results (last column). Interestingly, samples 5 and 10 appear to indicate RSV infection. Sample 18 also shows

a low presence of SARS-CoV-2 which is both below the threshold for calling a SARS-CoV-2 positive and maybe below the limit of detection of the PCR test.

Example 2: Low RNA Quality in Swab Samples

[0094] Several embodiments investigate RNA quality to achieve the desired RNA quality. Low quality RNA can result in low quality sequencing. Experiment with the ThermoFisher® Extraction Kit and the Illumina® Library Preparation Kit on 10 contrived samples resulted in low quality RNA in the samples from the sequencing metrics. Table 4 shows the number of reads in each sample before and after sequencing. 4 samples failed and the filtering removed anywhere from 40% - 75% of the reads for all but one sample due to RNA quality issues. Nothing is apparently different between the samples that succeeded and failed that can be observed from inspection of the physical sample which suggests that low RNA quality is an inherent property of nasal swab samples.

Table 4. Read counts before and after filtering for SwabSeq Metagenomic Diagnostic Platform

Sample	Reads Before Filtering	Reads After Filtering
1	106,974,802	26,547,804
2	83,046,300	25,067,986
3	53,482,892	48,179,934
4	42	12
5	152,453,294	98,608,552
6	6,581,514	1,353,444
7	178,606,460	114,643,644
8	0	0
9	12	12
10	216	80

[0095] Through both analysis of metrics and bioinformatics analysis, the source of the problem is identified. The low RNA quality after nucleic acid extraction can cause issues with the library preparation (Illumina® TruSeq) resulting in many adapter dimers. Figure 6 illustrates a histogram of bases per read cycle (read position) showing the adapter sequence in a highly enriched pattern.

[0096] Additional analysis looking for overrepresented sequences in the samples and observed several consistent patterns (such as sequences

CGAGCCCACGAGACACACAGGTGGA and CCTGTGTGTCTCGTGGGCTCGGAGA) that occurs hundreds of thousands of times in some of the samples are performed and is related to the adapter sequence. A likely explanation is that the samples had too little input nucleic acids which then resulted in adapter dimers.

[0097] To address this issue, additional RNA/DNA extraction kits experiments are tested out. The Zymo Research® Quick-DNA/RNA Viral MagBead showed superior performance. Some embodiments use a different library preparation protocol which is optimized for lower quality input. NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® (E6420L) is used.

[0098] Several embodiments perform experiments with depletion technologies including (but not limited to) robosomal deplation and CRISPR-based depletion (from Jumpcode). Depletion technologies can potentially improve the scalability and allow for more samples being processed. In most cases, depletion technologies further damage the poor RNA quality of the samples and perform worse than not using the technology. For this reason, some embodiments proceed without using depletion technology (with an exception below for using Jumpcode on the more scalable workflow).

[0099] Based on the experiments, about 4 million expected reads are needed to obtain enough reads for each sample. Some reads are lost at the sequencer level which do not pass sequencer level filters and additional reads are filtered by fastp. Finally, reads that contain artifacts are often not classified by kraken 2. Table 5 shows expected runtimes for different read configurations for the four Illumina® sequencers and their flowcell options where the total sequencing time is less than 15 hours. Other configurations with a sequencing time of greater than 15 hours are not practical to obtain a 24-hour turnaround. The sample capacity is estimated based on the number of expected reads for each flowcell.

Table 5. Expected runtimes for Illumina® Sequencer by Flowcell for a given read length.

Sequencer	Flow cell	Runtime	Sample Capacity	$ per Sample
MiniSeq	Rapid 50 bp	3.3 hours	5	$200
MiniSeq	Rapid 100 bp	5.1 hours	5	$200
NextSeq 550	High 75 bp	11 hours	100	$16
NextSeq 2000	P2 2 x 50 bp	13 hours	100	$14

NextSeq 2000	P3 1 x 50 bp	11 hours	300	$10
NovaSeq X Plus	10B 2 x 50 bp	18 hours	2,500	< $4
NovaSeq X Plus	25B 2 x 150 bp	N/A	6,500	< $2.5

[00100] Based on the experiments, several embodiments provide that: 1). The NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® paired with the Zymo Research Quick-DNA/RNA Viral MagBead resulted in good performance; 2). The NextSeq 550 and NextSeq 2000 that their performance is similar with the difference in sequencing depth (e.g., Table 4); 3). The NextSeq 550 and MiniSeq do not generate enough reads in order to process a full rack of samples and their use is not as practical from a logistical standpoint given the amount of effort it takes to prepare a library.

[00101] Building upon the experiments, several embodiments proceed with performing a final limit of detection and final clinical validation. A set of samples are prepared that can lead to a complete limit of detection with enough samples for both the preliminary limit of detection and the confirmatory limit of detection. 2 plates of contrived samples are prepared with 12 samples at each of the following concentrations:16,000 GCE/ml, 8,000 GCE/ml 4,000 GCE/ml, 2,000 GCE/ml, 1,000 GCE/ml, 500 GCE/ml, 250 GCE/ml, and 125 GCE/ml and a plate of 96 samples with 0 concentration. One contrived sample plate and one third of the 0 concentration plate were run on a NextSeq 2000 (128 samples) and both contrived plates and the 0-concentration plate on the NovaSeq 6000 (288 samples). The plan was to use the first 4 samples at a concentration for the preliminary limit of detection and the remaining 20 to perform the confirmation. This effort was successful and generated the data for the limit of detection for SARS-CoV-2 as described below.

[00102] Several embodiments obtain the data necessary to determine a final limit of detection for SARS-CoV-2 for SwabSeq Metagenomic Diagnostic Platform using the Zymo Quick-DNA/RNA Viral Magbead extraction and the NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® described in the next section. Previous experiments demonstrated that once the library is created, which sequencer is used only affects the total amount of reads and thus throughput of the assay. The analysis of clinical

samples obtains a clinical validation for the same assay which is described in the following section

Example 3: SwabSeq Metagenomic Diagnostic Platform Final Limit of Detection Results for SARS-CoV-2

[00103] Preliminary LoD and confirmatory LoD testing determines the lowest detectable concentration of SARS-CoV-2 at which approximately 95% of all (true positive) replicates tested positive. Using the contrived samples produced as described above, the preliminary LoD is defined as the lowest concentration where 3 of 4 replicates are identified correctly. The results of the preliminary LoD testing are summarized in Table 6.

Table 6. LoD Determination

Concentration tested (GCE/mL)	**Positive Replicates (n=4) (nasal swab samples)**
16000	4/4
8000	4/4
4000	4/4
2000	4/4
1000	4/4
500	2/4
250	0/4
125	2/4

[00104] LoD confirmation testing was performed by testing twenty (20) replicates at the preliminary LoD concentration determined above (1000 GCE/ml). Acceptance criteria for confirmation of the LoD was that at least 95% of the replicates (≥ 19/20) test positive. The LoD was confirmed to be 1000 GCE/mL for mid-turbinate samples run. Several embodiments also evaluated twenty (20) replicates at 0 concentration, and the LoD was confirmed to be 0 GCE/mL.

Example 4: SwabSeq Metagenomic Diagnostic Platform Clinical Validation for Detection of SARS-CoV-2

[00105] Clinical samples are remnant samples from COVID-19 testing submitted to the SwabSeq laboratory. Samples from individuals with symptoms consistent with COVID-19

submitted for COVID-19 testing and transported to the laboratory at 2 °C. After samples were accessioned but prior to samples being analyzed with the SwabSeq COVID-19 Diagnostic test, 100 µL is removed from each sample tube to be extracted using the SwabSeq Agnostic Diagnostic Test Extraction procedure. The SwabSeq COVID-19 Diagnostic test only requires no more than about 100µL of sample volume and a total of about 750 µL of sample is collected from each patient and thus the removal of the sample does not affect the results of the SwabSeq COVID-19 Diagnostic test. The resulting RNA extract is then stored at -80 °C which is consistent with recommendations for the NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® (E6420L). The advantages of this approach to obtain remnant samples is that the samples are extracted as fresh samples and not freeze/thawed samples, and the SwabSeq COVID-19 Diagnostic Test performs heat extraction which may damage nucleic acids.

[00106] Several embodiments perform a clinical validation using samples that are mid turbinate swab samples collected by a healthcare worker in 0.9% saline from individuals experiencing symptoms consistent with a respiratory infection. These samples were evaluated with a PCR assay. Remnant sample is then used for testing with the SwabSeq Metagenomic Diagnostic Platform. Results are summarized in Table 7. A total of 425 samples were included in the clinical validation.

Table 7. Evaluation with Clinical Specimens with 95% score confidence intervals

		PCR Comparator Assay		
		Detected	**Not Detected**	**Total**
SwabSeq Metagenomic Diagnostic Platform	**Detected**	50	4	54
	Not Detected	3	364	367
	Inconclusive	0	4	4
	Total	53	372	425
Positive Agreement		94.34% (50/53); 84.34% to 98.82%		
Negative Agreement		98.91% (364/368); 97.24% to 99.70%		
Overall Agreement		98.34% (414/421); 96.60% to 99.33%		

Example 5: SwabSeq Metagenomic Diagnostic Platform Bioinformatics Pipeline Development and Validation

[00107] Low sample quality and other factors combined with the use of public databases that may themselves have errors can potentially cause false positives. Several embodiments provide a new pipeline that is designed to minimize the possibility of false positives even when using databases that may contain errors. The idea is to determine in advance a set of curated genomes for the pathogen target and organisms that are likely present in the samples. Simulation using these curated genomes can be used to evaluate the pipeline to verify that even if there are inaccuracies in the public databases, the pipeline will be accurate with respect to the curated genomes. This approach reduces the possibility of false positives when using public databases.

[00108] In addition, if inaccuracies are observed in the simulations, certain embodiments can remedy them by identifying what parts of the genomes are ambiguous and filter reads out from those regions. Specifically, a "contaminant kmer list" which is used to filter the reads as part of the pipeline is obtained from these simulations. The procedure to obtain such a list is described above.

[00109] Several embodiments provide detailed experiments to obtain such a list. Using the curated genome for each target and commonly occurring organism in respiratory samples, 1,000,000 reads can be generated using wgsim from the curated organism and run through the pipeline. The prediction of genome of origin from the pipeline is described. If the pipeline is perfect, all 1,000,000 reads would be classified as that genome of origin. However, since reads are short and genomes have a certain amount of homology, reads will be misclassified. A set of incorrectly classified reads for each simulation can be identified. For a pathogen target, a read is considered misclassified if it is classified as human, bacteria or another pathogen target. For simulations from commonly present organism, we consider reads misclassified if they are classified as a pathogen target. From each simulation, kmers of length 31 are extracted from the reads and any kmers that occur at least 3 times in a "contaminant kmers list" can be added. The intuition is that these kmers are characteristic of regions of the genome which are inherently ambiguous as they lead to multiple misclassified reads. The pipeline filters out any read that contains

such a kmer. Table 8 shows this approach for 3 organisms, SARS-CoV-2, RSV and Haemophilus influenzae. For each organism, 1,000,000 reads were simulated and run through the pipeline. The first three columns show the results of the simulations. Entries in the table with an asterisk (*) are misclassified reads. Pathogen targets with 0 reads for these three organisms are omitted for clarity. The SARS-CoV-2 simulations resulted in 339 misclassified reads which then resulted in 51 kmers which were then added to the "contaminant kmer list." The RSV simulations added another 204 kmers to this list as well. The fourth column illustrates the effect of this approach by performing the same simulation and generating 1,000,000 reads for SARS-CoV-2 but filtering the reads that match a kmer. The results are that 4,492 reads were filtered out and the number of human and SARS reads were reduced compared to without filtering.

Table 8. Example of simulation studies of pathogens to generate a contaminant kmer list. 1,000,000 reads are simulated for each pathogen target and classified using the pipeline. Reads that are misclassified are noted by an asterisk (*). Kmers are extracted from these reads and any commonly occurring kmers (n>=3) are included in a contaminant list as they are characteristic of ambiguous genomic regions that can cause misclassified reads.

	Simulated Organism			
	SARS-CoV-2 (before filtering)	**RSV (before filtering)**	**Influzena A (before filtering)**	**SARS-CoV-2 (after filtering)**
Number of Simulated Reads	1,000,000	1,000,000	1,000,000	1,000,000
Human	79*	43*	7*	12*
Bacteria	73*	130*	66*	89*
SARS-CoV-2	939,195	0	0	937,863
Human Coronavirus HKU1	0	0	0	2*
Influenza A	0	0	941,923	0
RSV	0	805,674	0	0
Human Metapneumovirus	0	3*	0	0
Adenovirus	0	3*	0	0
SARS	187*	0	0	155*
Number of Reads Misclassified	339	179	73	258
# Kmers added to contaminant list	51	204	58	N/A
Reads filtered	0	0	0	4,492

[00110] Some embodiments provide validation of SwabSeq metagenomic bioinformatics pipeline. The following criteria for validating the pipeline to both correctly classify pathogen targets as well as be robust to false positives caused by other organisms present in the sample are developed. For a pathogen target to be validated from simulation studies, it is required that: 1). At least 50% of the simulated reads are correctly classified by the pipeline; 2). No more than 0.1% reads are misclassified to either Human, Bacteria or another pathogen target. In addition, it is required that from each organism commonly present in the samples: no more than 0.1% of the reads are misclassified as a pathogen target. Table 13 shows this validation for 3 pathogen targets. Without the additional contaminant kmer filtering, the pipeline is valid for those three targets. Table 9 shows the validation for 4 commonly occurring organisms in nasal samples.

Table 9. Example of simulation studies of common organisms in nasal samples. 1,000,000 reads are simulated for each common and classified using the pipeline. Reads that are classified as pathogen targets are considered misclassified. In these examples, no reads are misclassified.

	Simulated Organism			
	Human	**Staphylococcus epidermidis**	**Haemophilus influenzae**	**Streptococcus salivarius**
Number of Simulated Reads	1,000,000	1,000,000	1,000,000	1,000,000
Human	953,998	0	19	10
Bacteria	97	930,367	943,198	944,729
SARS-CoV-2	0	0	0	0
Human Coronavirus HKU1	0	0	0	0
Influenza A	0	0	0	0
RSV	0	0	0	0
Human Metapneumovirus	0	0	0	0
Adenovirus	0	0	0	0
SARS	0	0	0	0
Number of Reads Misclassified	0	0	0	0
# Kmers added to contaminant list	0	0	0	0
Reads filtered	0	0	0	0

[00111] Several embodiments show approaches to clean errors in the database by utilizing the technique above. Any unexpected results where reads from the simulated genomes match in other genomes suggest the possibility of the errors. Some embodiments developed manual approaches to evaluate these errors as well as automated approaches using machine learning or artificial intelligence (AI) techniques to identify database errors.

Example 6: Validate SwabSeq Agnostic Assay Using Contrived Samples from Multiple Pathogens

[00112] Several embodiments include a no-template control (NTC) to measure the amount of background nucleic acid detected in the assay and a positive control to verify the functioning of the assay. Some embodiments use "additional positive controls" in each sample that can be used to verify that the assay is working correctly. Certain embodiments include the Bacteriophage MS2 in each of the samples as an additional positive control.

[00113] A no-template control is the assay applied to the extraction buffer without a sample present. The purpose of a no-template control (NTC) is to determine the amount of background nucleic acids in the equipment and reagents that are detectable in the assay. The NTC can identify if the assay either has very high levels of contamination or is not functioning properly. After processing the assay, the NTC position should not detect any pathogens. In the assay, an NTC is implemented by including nuclease free water (NFW) in a well prior to extraction.

[00114] Some embodiments include at least one positive control in each run of the assay. There are two types of positive controls. The first is a sample which is a known positive verified by an alternative technology (such as a sample verified with the Roche ePlex or a saliva sample with a positive verified by the SwabSeq COVID-19 Diagnostic Platform (targeted PCR NGS test). For this type of control, the assay identified the correct pathogen for this sample. This positive control is used when performing validation experiments when measuring the performance of the assay against other assays. The second type of positive control is a contrived control which is used when applying the

assay to unknown samples. A mixture of multiple Twist pathogen controls is created and included on the plate. The mixture is varied in each control so that many of the target pathogens are included in some of the positive controls. For this type of control, the assay correctly identified all the components of the mixture.

[00115] Some embodiments include a known quantity of the Bacteriophage MS2 in each sample. The purpose of this control is to verify that the assay is working correctly by checking to see if enough reads from MS2 are observed as well as to quantify the amount of a pathogen if it's detected in the sample. Since a known quantity of MS2 is added to each sample, the ratio of the number of pathogens reads vs the number of MS2 reads is a robust measure of the amount of pathogen in the sample. Specifically, 10^6 PFU per mL of MS2 is added in each sample prior to extraction.

[00116] Some embodiments provide experiments validating inclusion of MS2 in assay. Specifically, there are three aspects of the assay related to MS2 and NTC: 1). How much MS2 should include in each sample; 2). What substance should be used for the NTC; 3). What threshold should be used for detection of a pathogen relative to the amount of MS2.

[00117] To address the first two aspects, the MS2 control and various options for the no-template control (NTC) are optimized on a sequencing run. Two quantities of MS2 (10^6 PFU/mL and 10^3 PFU/mL) are used. Both nuclease free water and DNA/RNA Shield are used as the NTC. Fresh saliva samples where MS2 is added. In the same run, unrelated aspects of the assay are optimized and included clinical saliva samples which were extracted previously without adding MS2 that have SwabSeq COVID-19 Diagnostic results (targeted NGS) to validate the saliva sample type. In addition, the entire set of experiments was replicated once for performing total nucleic acid extraction (DNA + RNA extract) and once for RNA extraction because the performance of total nucleic extraction may potentially identify DNA viruses. Table 10 shows the layout of this run. Given that the assay uses RNA extraction, the relevant columns of the plate are 7 and 8.

Table 10. Layout of plate for optimizing MS2 and NTC controls.

MS2	DNA + RNA extract						RNA extract					
	1	2	3	4	5	6	7	8	9	10	11	12

A	Saliva + MS2 at 10^6 PFU/mL	NFW+MS2 (10^6 PFU/mL) NTC	Clinical Saliva Samples	Saliva + MS2 at 10^6 PFU/mL	NFW+MS2 (10^6 PFU/mL)	Clinical Saliva Samples
B		NFW+MS2 (10^3 PFU/mL) NTC			NFW+MS2 (10^3 PFU/mL)	
C		DNA/RNA shield+MS2 (10^6 PFU/mL) NTC			DNA/RNA shield+MS2 (10^6 PFU/mL)	
D		DNA/RNA shield+MS2 (10^3 PFU/mL) NTC			DNA/RNA shield+MS2 (10^3 PFU/mL)	
E	Saliva + MS2 at 10^3 PFU/mL	NFW NTC		Saliva + MS2 at 10^3 PFU/mL	NFW NTC	
F						
G		DNA/RNA Shield NTC			DNA/RNA Shield NTC	
H						

[00118] Based on Table 10, MS2 above the threshold at certain positions and not at others are observed. Table 11 shows the relevant positions expected to observe MS2 and the positions where to observe MS2.

Table 11. MS2 positions in sequencing run (Table 10) corresponding to RNA extraction. Column 7 positions correspond to saliva samples with MS2 added prior to extraction. Column 8 corresponds to NTC controls. The number in Table 16 is the quantity of MS2 (PFU/mL) added to the sample. True positions show the positions expected to observe

S2 above the threshold. In the Actual column, the actual observed positions are bolded. One false positive position is marked with x.

True positions	MS2		Actual	MS2	
	7	8		7	8
A	**10^6**	**10^6**	A	**10^6**	**10^6**
B	**10^6**	**10^3**	B	**10^6**	**10^3**
C	**10^6**	**10^6**	C	**10^6**	**10^6**
D	**10^6**	**10^3**	D	**10^6**	**10^3**
E	**10^3**		E	**10^3**	x
F	**10^3**		F	**10^3**	
G	**10^3**		G	10^3	
H	**10^3**		H	10^3	

[00119] From the results of Table 11, MS2 is consistently detected in both clinical and NTC samples if added at 10^6 PFU/ml concentration and this is why 10^6 PFU/ml concentration of MS2 is used in the assay.

[00120] In addition, from this experiment, no detectable difference is observed between the two options of NTC which are nuclease free water (NFW in positions A08, B08, E08 and F08) and DNA/RNA shield (in positions C08, D08, G08 and H08) (data not shown). This is not unexpected since the first step of the RNA extraction protocol adds DNA/RNA shield. For this reason, nuclease free water is used as the NTC.

[00121] The third aspect of the assay related to MS2 is the threshold. The detection threshold for pathogens is the ratio of the number of pathogen reads to the number of MS2 reads. Since the same amount of MS2 is added to each sample, this threshold is much more robust and closer to an actual pathogen concentration than just the number of reads of the pathogen. The threshold of 0.01 which is high enough to filter out any contamination or other artifacts but low enough to identify all the clinical samples is used.

[00122] Several embodiments utilize contrived samples which were created using Twist Respiratory Virus Controls. Contrived samples are prepared by first obtaining negative samples from at least three individuals using mid turbinate swabs which were collected into 750 µl of 0.9% saline. Samples are then pooled together to create pooled negative nasal samples. Contrived samples are prepared by adding a quantity of the appropriate Twist Respiratory Virus Control to achieve a desired concentration. Positive samples were prepared at different concentrations typically 4000 GCE/ml, 2000 GCE/ml, 1000

GCE/ml, 500 GCE/ml, 250 GCE/ml and 125 GCE/ml, but in some cases in larger ranges. Table 12 shows the layout of the first LoD plate.

Table 12. Layout of Limit of Detection plate for Coronavirus OC43 and Enterovirus D68 pathogens.

	Virus	Contrived Concentration GCE/ml											
		1	2	3	4	5	6	7	8	9	10	11	12
A	Corona OC43	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
B	Corona OC43	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
C	Corona OC43	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
D	Corona OC43	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
E	Enterovirus	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
F	Enterovirus	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
G	Enterovirus	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31
H	Enterovirus	64,000	32,000	16,000	8,000	4,000	2,000	1,000	500	250	125	62	31

[00123] Over the 5 plates in this format, the following pathogens were examined: Human Coronavirus OC43, Human Coronavirus 229E, Human Coronavirus NL63, Enterovirus D68, Rhinovirus, Influenza B, Influenza A H1N1, Influenza A H3N2, Parainfluenza 1, Parainfluenza 4. The limit of detection is the lowest concentration where 2 out of 3 replicates detect the virus. 4 replicates were prepared when one of them is inconclusive. The estimated Limit of Detection for each of the samples is shown in Table 13.

Table 13. Estimated Limit of Detection for 10 pathogens using Twist Respiratory Virus Controls

Pathogen	Preliminary LoD
Human coronavirus OC43	1000 GCE/ml
Human coronavirus 229E	1000 GCE/ml
Human coronavirus NL63	1000 GCE/ml
Influenza A H1N1	2000 GCE/ml
Influenza A H3N2	1000 GCE/ml
Influenza B	1000 GCE/ml

Rhinovirus A	500 GCE/ml
Enterovirus D68	2000 GCE/ml
Parainfluenza 1	1000 GCE/ml
Parainfluenza 4	1000 GCE/ml

[00124] In Table 13, the LoD is already comparable to multiplex-PCR tests such as the Roche ePlex. Because the ranges of limit of detection is in the same range for all pathogens, after modifying the assay and performing new LoD experiments, only the concentrations 4000 GCE/ml, 2000 GCE/ml, 1000 GCE/ml, 500 GCE/ml, 250 GCE/ml and 125 GCE/ml are considered for the experiments and can put 4 pathogens in each plate. The plate is shown in Table 14 which contains Rhinovirus, Influenza H1N1, Mumps and Parainfluenza 1.

Table 14. Layout of Limit of Detection (LoD) plate for Rhinovirus, Mumps, Influenza H1N1 and Parainfluenza 1 pathogens

	Virus 1	Contrived Concentration GCE/ml												Virus 2
		1	2	3	4	5	6	7	8	9	10	11	12	
A	Rhinovirus	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Influenza H1N1
B	Rhinovirus	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Influenza H1N1
C	Rhinovirus	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Influenza H1N1
D	Rhinovirus	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Influenza H1N1
E	Mumps	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Parainfluenza 1
F	Mumps	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Parainfluenza 1
G	Mumps	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Parainfluenza 1
H	Mumps	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Parainfluenza 1

[00125] In addition, the plate shown in Table 14, 3 more plates were used to perform the limit of detection studies. The plate contained Enterovirus D68 and Influenza B. The

plate contained Parainfluenza 4, Human Coronavirus NL63, Human Coronavirus 229E and Measles. The plate contained Influenza H3N2 and Human Coronavirus OC43.

[00126] Over the four plates that were run during this period, limit of detection for the following pathogens: Enterovirus D68, Rhinovirus, Influenza A H1N1, Influenza A H3N2, Influenza B, Human Coronavirus NL63, Human Coronavirus 229E, Human Coronavirus OC43, Parainfluenza 1, Parainfluenza 4, Mumps, Measles were estimated. Influenza A H3N2 and Human Coronavirus OC43 were not immediately available so Mumps and Measles were added.

[00127] Surprisingly, the results from the second round of Limit of Detection studies are that the assay detected all the samples resulting in a limit of detection of 125 GCE/ml for all pathogens which is an improvement of 8x or 4x in the LoD. There are several reasons why the limit of detection improved. First, incorporating automation results in better performing libraries. The improved libraries obtained on average 4.8M reads per sample that pass filters compared to 3.1M reads. In addition, the sequencing quality of the sample has improved based on QC metrics and many more of the reads in the newer samples are informative for identifying viruses. Finally, adding the MS2 control sets the threshold relative to the control rather than the total reads which also improves the performance.

[00128] However, there are some caveats to this analysis. First, Twist provides an approximate concentration for their materials. In addition, the Twist controls are synthetically generated as opposed to heat inactivated. Thus, it is also possible that the improvement is specific to the batch of controls compared to different controls.

[00129] NGS technology enables one to directly quantify the amount of carryover from previous runs. The LoD plates are utilized to conduct a carryover study. The two LoD plates contain very high viral loads of distinct pathogens. They were both generated by the same operator using the same equipment and run 2 days apart. The first plate has a layout shown in Table 19. The second plate has a layout shown in Table 15.

Table 15. Layout of Limit of Detection (LoD) for Parainfluenza 4, Measles, Coronavirus NL63 and Coronavirus 229E pathogens

	Virus 1	Contrived Concentration GCE/ml												Virus 2
		1	2	3	4	5	6	7	8	9	10	11	12	

A	Parainfluenza 4	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus NL63
B	Parainfluenza 4	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus NL63
C	Parainfluenza 4	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus NL63
D	Parainfluenza 4	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus NL63
E	Measles	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus 229E
F	Measles	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus 229E
G	Measles	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus 229E
H	Measles	4,000	2,000	1,000	500	250	125	4,000	2,000	1,000	500	250	125	Coronavirus 229E

[00130] Rhinovirus, Mumps, Influenza H1N1 and Parainfluenza 1 are present. These two plates allow the direct measurement of carryover by quantifying the number of reads corresponding to pathogens from the first plate detected on the second plate. Table 16 shows the number of reads of pathogens from the first plate on the second plate which is evidence of carryover. The carryover is not in the same wells as the samples were in the previous plate and this is expected as the carryover can occur because of contamination from any of the processes of processing the samples.

Table 16. Pathogen reads on the second LoD plate due to carryover from the first plate

Well	**Pathogen**	**Number of Reads**
A02	Mumps	12
A07	Mumps	16
A08	Mumps	29
A10	Mumps	23
B01	Rhinovirus	30
B11	Influenza A H1N1	12
D11	Mumps	13

[00131] The amount of carryover can be estimated by computing the ratio of the number of reads observed for each pathogen in the second plate divided by the number of reads observed in the first plate. Table 17 shows the results for each pathogen and an overall estimate. Carryover from over 275 million pathogen reads in the second plate only

generated 122 pathogen reads in the first plate. Actual plates of clinical samples will contain far fewer pathogen reads (typically under 1 million reads) and thus the amount of carryover will be even smaller than here. If assuming 1 million pathogen samples in a previous run and the same carryover rate estimated here, the amount of carryover would be in the range of 1-2 reads. The carryover contamination rate is much smaller than other sources of errors and not likely a contributor to false positives.

Table 17. Estimate of carryover per pathogen and total for plates run on the first and the second plates

Pathogen	Number of Reads on the second plate	Number of Reads on the first plate	Estimated Carryover Rate
Mumps	80	123,169,004	6.50 * 10^-7
Rhinovirus	30	127,064,815	2.36 * 10^-7
Influenza A H1N1	12	20,331,374	5.90 * 10^-7
Parainfluenza 1	0	4,957,386	0
Total 3/26/24 Pathogens	**122**	**275,522,579**	**4.43 * 10^-7**

[00132] Cross-contamination is a fundamental problem for NGS Diagnostic technologies. The rate of cross-contamination can be estimated by utilizing samples which are known positives for pathogens. In one plate, there are 5 samples which were confirmed to be positive for Parainfluenza 3 using a Roche ePlex Respiratory Panel comparator. For the 5 known Parainfluenza 3 positives, a dramatic range of the number of pathogen reads for each positive is shown in Table 18.

Table 18. Number of Parainfluenza 3 reads from known positive samples in the plate

Well	Number of Parainfluenza 3 Reads
B1	199,585
B3	22,039
E1	1,812,069
F1	1,053,167
H1	16

[00133] From Table 18, even small cross-contamination from a very high positive (such as from well E1), would create a signal stronger than a low positive (such as well H1). Thus cross-contamination, even in very small amounts, can cause false positives.

[00134] Figures 6A through 6C show the results of analysis of the plate both using the manual library preparation protocol and the automated library preparation protocol. In the library preparation manual there is a much higher level of cross contamination compared to the automated library preparation. Figure 6A shows 5 True Positive positions of Parainfluenza 3. Figure 6B shows predicted positions using manual library preparation protocol where blue are true positive predictions and red are false positives due to cross-contamination. Figure 6C shows predicted predictions using automated library reparation protocol. Positions are highlighted only if there are more than 10 reads present.

[00135] For both library preparation approaches, the amount of contamination can be quantified by considering the reads in neighboring well as a percentage of the source reads. Table 19 shows the number of reads for this run for the Parainfluenza 3 pathogen.

Table 19. Cross-contamination analysis for true positives of Parainfluenza 3 pathogen. Each row shows the position of the true position which is the potential source of cross-contamination, the neighboring position and the number of reads for both the source and neighboring positions for manual and automated library preparation protocols. The rate is estimated by the ratio of the source reads to neighbor reads. Only positions where cross-contamination occurred are shown in the table.

True Positive Position (Source)	Neighboring Position	Manual Library Prep Source Reads	Manual Library Prep Cross-contamination Reads (Rate)	Automated Library Prep Source Reads	Automated Library Prep Cross-contamination Reads (Rate)
E1	D1	639,322	876 (0.18%)	1,812,069	0 (0%)
E1	D2	639,322	53,489 (8.37%)	1,812,069	0 (0%)
E1	E2	639,322	73 (0.01%)	1,812,069	0 (0%)
E1	E3	639,322	1,021 (0.16%)	1,812,069	0 (0%)
F1	G1	1,270,046	11,724 (1.84%)	1,053,167	0 (0%)

[00136] A precision study is performed by applying the assay on the same set of samples twice using a different sequencer. One run was sequenced on a NextSeq 2000.

The same set of samples was repeated in the assay including new RNA extraction, library preparation and sequencing. The second run utilized a NovaSeq X Plus. There was a total of 54 positive samples for various pathogens on the first run and 60 positive samples on the second run. Table 20 shows the concordance of positives between runs.

Table 20. Repeatability (Precision) results for two runs of clinical samples. The Enterovirus positives on the second run are marked with an asterisk because they are lower positives and only detectable at a deeper sequencing depth than the first run.

Pathogen	Both Runs	First Run	Second Run
Adenovirus	7	0	0
Enterovirus/Rhinovirus	10	0	6*
Influenza A	8	0	0
RSV	28	1	1

[00137] The only pathogen where there is a difference is for Enterovirus where there are 6 additional positives on the second run. In fact, the motivation for repeating these samples was the hypothesis that there are potentially additional low positive Enterovirus samples on the plate. The NovaSeq X Plus generates substantially more sequence and the median sequence depth was 20x higher on the second run which explains why additional positives were identified. Putting these samples aside, the experiment shows that the assay is highly reproducible.

[00138] To establish analytical specificity, contrived samples which contain mixtures of pathogens using material obtained from Twist where each pathogen has a concentration of 2x of the estimated LoD are generated. The assay is run on these samples and verifies that each pathogen is identified in the sample. Table 21 shows the design of this experiment.

Table 21. Design of Analytical Specificity Experiment. The plate contained 18 samples contrived samples, each with three pathogens that were created using Twist materials. 9 different combinations were used with 2 samples per combination. The exact combinations of pathogens in each sample are shown in the table. The experiment evaluated combinations of similar pathogens (F9, F10), combinations of different

pathogens (F11, F12, G7, G8, G11, G12, H7, H8, H9, H10) and combinations of two similar and one different pathogen (F7, F8, G9, G10, H11, H12).

Specificity	Sample Well								
Pathogen	**F7, F8**	**F9, F10**	**F11, F12**	**G7, G8**	**G9, G10**	**G11, G12**	**H7, H8**	**H9, H10**	**H11, H12**
Parainfluenza 1			X			X	X		X
Parainfluenza 4			X					X	
Influenza H1N1	X			X	X		X		
Influenza H3N2	X		X		X				
Coronavirus OC43		X							X
Coronavirus 229E		X							X
Coronavirus NL63		X							
Mumps				X	X	X	X		
Measles	X			X				X	
Rhinovirus						X		X	

[00139] Table 22 shows the results of three representative samples F9, F7 and H9 of three similar pathogens, two similar and one different and three different pathogens respectively. As shown, the assay can identify the correct pathogens within the mixtures. The assay correctly identified the pathogens in each of the 18 samples.

Table 22. Results of Analytical Specificity Samples. Results of three samples from Figure 12. As shown, the assay correctly identifies the components of the mixture of pathogens.

Well	**Identified Pathogen**	**Number of Reads**
F7	Influenza H1N1	1,715
F7	Influenza H3N2	2,301
F7	Measles	1,978
F9	Human Coronavirus 229E	156
F9	Human Coronavirus NL63	6,885
F9	Human Coronavirus OC43	2,441
H9	Rhinovirus	34,823
H9	Measles	1,772
H9	Parainfluenza 4	962

[00140] Some embodiments perform an interfering substances study considering 5 substances that can potentially interfere with the assay: Cough Drops, Mouth Wash, Cough Syrup, Bovine Mucin, Sore Throat Spray. For each substance, 3 contrived positive samples at 2x the LoD were generated and added to the substance. The substance was added to 3 negative samples. The assay was applied to the contrived samples. Results are shown in Table 23. None of the substances interfered with the assay as expected.

Table 23. Interfering Substances. Positives are contrived samples that were generated at 2x LoD prior to adding the substance. *Positive generated from Twist Influenza H3N2. **Positive generated from Twist Coronavirus OC43. The number in the table is the number of samples where the assay made a correct prediction.

Substance	Concentration	Positives	Negatives
Sore throat and cough lozenges (benzocaine/menthol**)**	3 mg/mL	3/3*	3/3
Mouth Wash	5% v/v	3/3*	3/3
Cough Syrup (e.g. Robitussin)	5%	3/3*	3/3
Bovine Mucin	2.5 mg/mL	3/3*	3/3
Clorospeptic Sore Throat spray	5% v/v	3/3**	3/3

[00141] Some embodiments perform a stability study using saliva samples to evaluate how consistent the assay results are by varying the time between sample collection and processing. The first step of the Zymo Quick-DNA/RNA Viral Magbead is to add the RNA/DNA Shield reagent. It is well established that RNA/DNA Shield preserves RNA for long periods of time stored at ambient temperature and it is well known that RNA analysis on samples which contain RNA/DNA Shield are consistent irrespective of how long analysis was performed after the RNA/DNA Shield was added. Thus, the key question in the stability study is to understand the effect on the assay of how soon after sample collection is RNA/DNA Shield added.

[00142] Some embodiments perform a stability study using saliva samples submitted for COVID-19 testing. Prior to COVID-19 testing, 3 aliquots were obtained from 96 samples. RNA/DNA Shield was added to one set of aliquots right away. These samples are referred to as the “Immediate” samples. RNA/DNA Shield was added to the second

aliquot after 3 to 6 days stored at room temperature. These samples are referred to "Delayed" samples. A third set of aliquots were frozen and thawed prior to RNA/DNA shield being added. These samples are referred to as "Frozen" samples. All three sets of 96 samples (288 samples total) had their libraries constructed together and sequenced together in order to minimize any additional sources of variance outside the stability study.

[00143] Table 24 shows the stability study results. Each entry shows a pathogen identified in the immediate samples, and the corresponding pathogen identified in the delayed samples. Both the number of reads and MS2 ratio are provided.

Table 24. Interference Study Results. The same samples were either treated with RNA/DNA Shield immediately upon arriving at the laboratory or delayed several days before being treated with RNA/DNA Shield simulating a longer time between collection and arrival in the laboratory. Each pathogen identified in the samples immediately analyzed was also identified in the delayed samples

		Immediate Samples		**Delayed (by Days) Samples**		
Well	Pathogen	Reads	MS2 Ratio	Days	Reads	MS2 Ratio
A07	Human Coronavirus HKU1	21	0.05455	5	57	0.07179
A12	SARS-CoV-2	287	0.05138	6	50	0.02846
C10	Rhinovirus	411	0.04099	6	377	0.05305
D10	Enterovirus	31	0.02160	6	18	0.01741
E02	Human Coronavirus 229E	744	0.18931	3	467	0.13895
F07	SARS-CoV-2	97,383	0.70583	5	7,701	1.44539
H06	HSV	19	0.01114	5	22	0.03767

[00144] Some embodiments use the LoD samples to demonstrate the accuracy of the bioinformatics pipeline. Only the contrived pathogens present in the plate to be identified by the pipeline. Specifically, the LoD plates contained the following pathogens: First plate contains Enterovirus D68 and Influenza B; second plate contains Parainfluenza 1, Influnenza H1N1, Mumps and Rhinovirus; third plate contains Parainfluenza 4, Human Coronavirus NL63, Human Coronavirus 229E and Measles; fourth plate contains Influenza H3N2 and Human Coronavirus OC43.

[00145] Some embodiments measure the accuracy of the pipeline by computing the number of misclassified reads or reads that do not match the contrived pathogens in each run. There are 0 misclassified reads for the first plate. For the second plate, the only

reads that are misclassified are some Influenza H1N1 reads which are classified as Influenza H3N2. However, the number of reads is below the 5% threshold which is used to identify pathogens within pathogen groups so will have no effect on the overall predictions. For the third plate, the only misclassified reads are the 122 reads in Table 25, or the reads classified to Parainfluenza 4, Measles, Coronavirus NL63 and Coronavirus 229E pathogens which is likely due to carryover rather than a bioinformatics pipeline error. No analysis was done on the fourth plate since part of that run contained mixtures of different pathogens and every pathogen was present in that plate. Table 25 summarizes these results.

Table 25. Misclassified reads from LoD studies.

Plate	Pathogens	Misclassified Reads
First	Entrovirus D68, Influenza B	None
Second	Parainfluenza 1, Influenza H1N1, Mumps, Rhinovirus	Only <5% Influenza H1N1 classified as Influenza H3N2
Third	Parainfluenza 4, Human Coronavirus NL63, Human Coronavirus 229E, Measles	122 reads from carryover (Table 21)

Example 7: Validate SwabSeq Agnostic Assay using clinical samples from multiple pathogens

[00146] The clinical validation of the SwabSeq Metagenomic Diagnostic Platform was performed over 3 sets of samples: 1). Remnant mid-turbinate swabs collected for COVID-19 testing by the SwabSeq Laboratory; 2). Remnant nasal pharyngeal swab samples; 3). Remnant saliva samples collected for COVID-19 testing. Table 26 shows a summary of the sets of samples.

Table 26. Summary of sets of samples in clinical validation.

Sample Type	Study Type	Comparator	Number of Samples	Number of Unique Samples
Mid-turbinate swab	Prospective	SwabSeq COVID-19	482	482
Nasopharyngeal swab*	Retrospective	Roche ePlex	1,160	460
Saliva	Prospective	SwabSeq COVID-19	1,248	1,056
Total			**2,890**	**1,998**

[00147] Of the three sets of samples, the second-row samples utilize a comparator that can detect multiple pathogens. The other two sets utilize a comparator that can only detect SARS-CoV-2 but have the advantage that they are prospective studies in that they are a set of samples randomly selected from samples submitted for COVID-19 testing. After the analytical validations on multiple pathogens and the clinical validation, the results of detecting pathogens other than SARS-CoV-2 in these two other sets. The samples include three different sample types (Mid-turbinate swab, nasal pharyngeal swab and saliva).

[00148] Some embodiments perform a clinical validation. A total of 460 remnant samples which were selected based on the results of analysis by the Roche ePlex Respiratory Pathogen Panel 2 were collected. 210 samples were positive for a pathogen on the ePlex panel and 250 were negative. Most of the samples were nasopharyngeal swabs, but other sample types were also included in the set of samples. The sample type distribution is shown in Table 27.

Table 27. Sample type distribution of set of samples

Sample Type	Count
Respiratory, Upper, Nasopharyngeal	378
Respiratory, Lower, Bronchoalveolar Lavage	49
Respiratory, Lower, Sputum-Expectorated	10
Respiratory, Upper, Nasal Washing	4
Respiratory, Upper, (Asst) Mid-turbinate	4
Respiratory, Lower, Tracheal Aspirate	4
Respiratory, Lower, Endotracheal	4
Respiratory, Upper, (Self) Mid-turbinate	1
Respiratory, Lower, Bronchoalveolar Lavage, Right	1
Respiratory, Upper, Nasopharyngeal Wash	1
Respiratory, Lower, Nasotracheal Suction	1
Respiratory, Lower, Other, Enter source information	1
Respiratory, Lower, Bronchoalveolar Lavage, Left	1
Respiratory, Lower, Sputum-Induced	1

[00149] The results of the comparison of the Roche ePlex and the SwabSeq Metagenomics Diagnostic Platform are in Table 28.

Table 28. Results of positives from samples. 210 positive samples and 250 negative samples were obtained, and the table shows results by pathogen for the 204 samples that passed QC and were analyzed. *The process of selecting samples for analysis may have included samples that are positive for either SARS-CoV-2 or Influenza A as negative samples. In analysis of the negative samples, only pathogens other than SARS-CoV-2 and Influenza A are considered.

Clinical Validation (Pathogen)	SwabSeq Metagenomic
Human Coronavirus	8 / 8
Human Metapneumovirus	6/ 6
Parainfluenza Virus 1	6 / 7
Parainfluenza Virus 2	5 / 6
Parainfluenza Virus 3	7 / 7
Parainfluenza Virus 4	5 / 6
Rhinovirus/ Enterovirus	27 / 34
RSV	40 / 40
Adenovirus	9 / 11
Influenza A	35 / 39
SARS-CoV-2	38 / 40
Negative	248/250 *

[00150] Except for Rhinovirus/Enterovirus, the results of the SwabSeq Metagenomic Diagnostic Platform are comparable to the Roche ePlex Respiratory Panel.

[00151] In the clinical validation of the assay to detect SARS-CoV-2, 482 mid-turbinate swab clinical samples were analyzed. The primary focus of the analysis was to estimate the concordance with the SwabSeq COVID-19 targeted NGS PCR test for detecting SARS-CoV-2. In these samples, many other pathogens other than SARS-CoV-2 were observed. Table 29 shows the pathogens other than SARS-CoV-2 that were detected in these samples. Since the comparator only detects SARS-CoV-2, it was unable to confirm

the presence or absence of any other pathogens. Based on the analytical and clinical studies for other pathogens, these pathogens are in fact present in the samples.

Table 29. Pathogens other than SARS-CoV-2 in SwabSeq Swab Samples

Pathogen	Count
Enterovirus	9
Human alphaherpesvirus 1 (HSV)	4
Coronavirus HKU1	1
Coronavirus NL63	1
Coronavirus OC43	1
Human orthopneumovirus (RSV)	11
Influenza A H1N1	3
Influenza A H3N2	1
Parainfluenza 3	1

[00152] Saliva as a sample type has many advantages, yet until the COVID-19 pandemic, saliva was not regularly utilized for infectious disease diagnostics. Many embodiments utilize the SwabSeq Metagenomic Diagnostic Platform to demonstrate that many pathogens can be detected in the saliva sample type. About 1,152 samples were analyzed. Table 30 shows the pathogens other than SARS-Cov-2 identified in the samples.

Table 30. Pathogens other than SARS-CoV-2 in SwabSeq Saliva Samples

Pathogen	Count
Enterovirus	6
Human alphaherpesvirus 1 (HSV)	3
Coronavirus HKU1	6
Coronavirus NL63	1
Coronavirus OC43	9
Coronavirus 229E	3
Human orthopneumovirus (RSV)	3
Influenza A H3N2	1

Example 8: Integrate the Assays into Automation

[00153] Automation can enable processing many samples simultaneously. Specifically, a plate of 96 samples or multiple plates of 96 sample can be run at the same time. Each step is designed to process one plate at a time. Two sequencing protocols: one protocol

uses the NextSeq 2000 P2 flowcell that can sequence 96 samples and one protocol uses the NextSeq 2000 P3 flowcell that can sequence 3 sets of 96 samples (288 total samples).

[00154] Some embodiments provide accessioning and nucleic acid extraction automation. Nucleic acid extraction has been performed using the Thermo Scientific® KingFisher Apex system. The extraction kit from Zymo Research® is compatible with this system. The output of the Kingfisher is a plate of the RNA extracts of the 96 input samples.

[00155] Since this is the first step of the assay, there are several workflow challenges related to sample tracking and accessioning that needed to be addressed. Part of the clinical validation was performed using clinical samples submitted for COVID-19 testing using the SwabSeq COVID-19 Diagnostic test which is a targeted NGS diagnostic. This diagnostic test does not perform nucleic acid extraction. In order to incorporate the SwabSeq Metagenomic Diagnostic Platform, a workflow that would perform extraction on the submitted samples is adapted. Racks of 96 samples are received in Thermo Scientific® Matrix 1.0mL tubes which each have three barcodes: a human readable barcode, a linear barcode on the side of the tube and a 2D barcode on the bottom of the tube. The racks are scanned using a barcode scanner which captures the barcodes and positions of each tube in the rack. 250 µL are aliquoted from each sample to a 96 well plate and the captured barcodes and positions are stored long with the plate barcode in the SwabSeq Laboratory Information Management System (LIMS). This plate can then be processed in KingFisher Apex system for extraction while the original samples can follow the SwabSeq COVID-19 Diagnostic test workflow. This process can extract nucleic acids from samples.

[00156] Some embodiments provide sequencing library preparation. In order to automate the sequencing library preparation, the specific library preparation protocol such as NEBNext® Single Cell/Low Input RNA Library Prep Kit can be used. To automate the assay, the first step is to modify the existing Biomek i7s to support the protocol. Cold blocks, magnetic modules and integration kits are used. A Thermocycler is not used on the Biomek which allows for a second plate to be processed while the first plate is in the thermocycler. The assay has 6 major steps, and, in each step, there is a combination of user time to set up the Biomek deck and add reagents, thermocycler incubations and Biomek automation. Table 31 shows an overview of the automated protocol steps. The

total time of the library preparation assay is approximately 9 hours for 1 plate. If a 2nd plate is processed at the same time, most of the steps can be completed while the first plate is in the thermocycler. The only exception is the cDNA cleanup which performs multiple incubations on the Biomek and repeatedly uses the magnetic module. Overall, as shown in the table, a second 96 well plate requires an additional 2 hours.

Table 31. Summary of Biomek automation for sequencing library preparation. There are 6 steps for the protocol and (almost) each step involves user preparation of the Biomek and reagents, Biomek actions and thermocycler incubations. Most of the user actions can occur during incubations and do not add any time to the protocol. Since the thermocycler is not on the Biomek deck, a second plate can be processed while the first is in the thermocycler.

Protocol Step	User Actions	Biomek Actions	Thermocycler Incubations
Reverse Transcription and Template Switching	1. Deck setup 2. Prepare RT master mix	1. Aliquot primers to Plate 2. Mix master mix to plate	5 minutes 100 minutes
cDNA amplification by PCR	1. Deck setup 2. Prepare reagents	1. Aliquot cDNA master mix into plate	70 minutes
cDNA cleanup	1. Deck setup	1. Aliquot beads 2. Multiple on deck incubations on and off magnets	None
Fragmentation and End Prep	1. Deck setup 2. Prepare reagents	1. Aliquot fragmentation master mix into plate 2. Mix clean cDNA with master mix	56 minutes
Adapter Ligation and clean up	1. Deck setup 2. Prepare reagents	1. Aliquot ligation master mix and adaptor 2. Mix with fragmentation 3. Perform bead cleanup	15 minutes 15 minutes
Library PCR and clean up	1. Deck setup 2. Prepare reagents	1. Stamp primer plate 2. Add master mix	30 minutes

[00157] There is one step in the library preparation which is not automated on the Biomek i7 which is "Quantification, Normalization and Pooling". In this step, each resulting library is quantified using a Qubit and all libraries are pooled together into a single tube at a proportion so that the resulting sequencing will have a similar number of total reads for each sample. This step is automated using other liquid handles.

[00158] An alternative approach to the Biomek i7 is a manual library preparation for a small number of samples which also requires a limited amount of hands on time. The

manual library preparation for 5 samples can be performed in approximately 6 hours. 5 samples are the number of samples that can be run on the MiniSeq Rapid kit. This process can also be automated on a smaller automated liquid handler such as an Opentrons OT-2. This process can also be automated up to 8 samples on a microfluidic liquid handler. The process can be automated for many more than 8 samples on a microfluidic liquid handler as microfluidic liquid handlers increase in their capacity.

[00159] Table 32 shows the runtimes of relevant Illumina flow cells for the final configurations. In the table, the three flow cells (MiniSeq Rapid 50 bp, NextSeq 2000 P2 2 x 50 bp and NextSeq P3 1 x 50 bp) are listed. The MiniSeq Rapid 50 bp and NextSeq P3 1 x 50 bp flow cells only collect 50 bp per read but this is enough for our application.

Table 32. Sequencing runtimes of Illumina flow cells relevant to our assay configurations. Flow cells marked with an * are flow cells which are utilized in the configurations in accordance with embodiments. ** The runtime for the MiniSeq Rapid assumes only single indices which is appropriate since we only have 5 samples in a MiniSeq Rapid run.

Illumina Flow Cell	**Runtime**
MiniSeq Rapid 50 bp*	3.3 hours**
MiniSeq Rapid 100 bp	5.1 hours**
NextSeq 2000 P2 2 x 50 bp*	13 hours
NextSeq 2000 P3 1 x 50 bp*	11 hours
NextSeq 2000 P3 2 x 50 bp	19 hours

[00160] In practice, the MiniSeq Rapid flow cell can generate 100 bases instead of 50 and the NextSeq P3 2 x 50 bp is practically the same cost as the NextSeq 2000 P3 1 x 50 bp. The only downside of these two flow cells is that they have longer runtime. Given that collecting this additional data was effectively free, the full 100 bp from the MiniSeq Rapid and NextSeq 2000 P3 2 x 50 bp flow cells were collected.

[00161] Collecting 50 bp instead of 100 bp can achieve the same results. Several embodiments reanalyzed the data using only 50 bp of each read instead of the full 100 bp. Table 33 shows the concordance of positives when analyzing 50 bp vs 100 bp of

each read. All positives are fully concordant showing that only 50 bp are necessary to identify viral pathogens.

Table 33. Concordance of analysis of 50 bp vs 100 bp showing that only 50 bp are necessary to identify viral pathogens.

Pathogen	Positive both in 50 bp and 100 bp data	Positive only in 100 bp data	Positives only in 50 bp data
Parainfluenza 1	6	0	0
Parainfluenza 2	6	0	0
Parainfluenza 3	5	0	0
Parainfluenza 4	5	0	0
Human Coronavirus	6	0	0
Rhinovirus/Enterovirus	4	0	0
Human Metapneumovirus	6	0	0

[00162] Bioinformatics Times. The SwabSeq Metagenomic Bioinformatics Pipeline has three contributors to its runtime. The first is the time it takes to get the data from the sequencer to the cloud for analysis. The second is to convert the data from BCL format to individual sample files. The third is the analysis of each sample file. The file transfer and BCL convert must be performed before any of the other analyses and the individual analyses can be performed in parallel. Table 34 shows the run time of each of these steps for each of our configurations.

Table 34. Length of time necessary for the bioinformatics pipeline to analyze the data generated by each configuration. For the two NextSeq configurations, after BCL conversion, sample processing is parallelized on Amazon Web Services.

Bioinformatics Step	MiniSeq Rapid	NextSeq P2	NextSeq P3
Number of Samples	5	96	192
File transfer	15 seconds	95 seconds	304 seconds
BCL Conversion	7 seconds	109 seconds	368 seconds
Number Sample Processing Nodes	1	16	32
Sample Processing Time	8 minutes	10 minutes	10 minutes
Total Time	< 10 minutes	14 minutes	22 minutes

[00163] Sample to Answer Time. For each step of the assay which include: nucleic acid extraction; the 6 steps of sequencing library preparation; quantification, normalization and pooling; sequencing; and bioinformatics. The time of each step can be decomposed into two parts, a preparation time and a sample to answer time contribution. The preparation step involves preparing reagents and equipment for the step and can be performed either before analysis starts or during the previous step. The sample to answer time depends on the number of samples being processed and is summarized in Table 35. The sample to answer for 3 different configurations: 96 samples, 192 samples and 5 samples.

Table 35. Sample to Answer Time for the three different configurations. Each step is broken down into preparatory time which are parts of the protocol that can be performed either before the samples are analyzed or during the previous step.

Sample to Answer Times	MiniSeq Rapid		NextSeq 2000 96		NextSeq 2000 192	
Flow Cell	MiniSeq Rapid 50 bp		NextSeq 2000 P2 2x50 bp		NextSeq 2000 P3 50 bp	
Number of Samples	5		96		192	
Times in hours	Preparation Time	Sample to Answer Time	Preparation Time	Sample to Answer Time	Preparation Time	Sample to Answer Time
Nucleic Acid Extraction	0.5	1	0.5	1	0.5	1
Reverse Transcription and Template Switching	0	1.9	0.2	1.9	0.2	2.2
cDNA amplification by PCR	0	1.1	0.3	1.1	0.4	1.1
cDNA cleanup	0	0.5	0.3	1.7	0.4	3.4
Fragmentation and End Prep	0	1	0.2	1	0.3	2
Adapter Ligation and clean up	0	1.1	0.3	1.1	0.4	2.2
Library PCR and clean up	0	0.2	0.2	1	0.3	1.1
Quantification, Normalization, Pooling, Cleanup	0	0.2	0.3	2	0.4	2.5
Sequencing	0.5	3.3	0.5	13	0.5	11
Bioinformatics	0	0.2	0	0.3	0	0.4
Total Sample to Answer Time		10.5		24.1		26.9

[00164] The third configuration processes 2 sets of 96 sample plates on a single Biomek i7. We chose to not install a thermocycler on the Biomek and instead use an external thermocycler so that a second plate can be prepared during the Biomek while the first is in the thermocycler. There are several constraints in automation. One constraint is that the cDNA cleanup step is performed on the Biomek deck and has many room temperature incubations. Since the deck is occupied throughout the step, it is impossible to overlap the two plates. A second constraint is that the result of the Fragmentation and End Prep step is not stable, and the Adapter Ligation and cleanup step must be started immediately afterwards. The simplest way to address this constraint is to perform both steps for one plate before performing both plates for the next plate. Figures 7A and 7B

show diagrams of how two plates are analyzed by single Biomek. We note that both plates are extracted simultaneously on different KingFisher systems.

[00165] Figures 7A and 7B illustrate interleaving of 2 96 sample plates on 1 Biomek i7. Most steps have a long thermocycler incubation at the end of the step which allows a second plate to be processed on the Biomek during the incubation. However, cDNA cleanup (L3) has no off-deck incubation step and the plates must complete the full step before releasing the deck. In addition, Fragmentation and End Prep (L4) is not stable so much be immediately followed by Adapter Ligation and cleanup (L5). Figure 7A shows current automation workflow. The extra time for the second plate is a little bit of time at the beginning and then the doubling of the L3, L4 and L5 steps which are performed serially. Figure 7B shows automation workflow that is optimized. L4 and L5 are interleaved since L4 ends with a long thermocycler incubation and L5 for the other plate can complete during that time.

[00166] There are several places that can be improved in the automated workflow. One step that is currently manual is the Quantification, Normalization and Pooling. This workflow can be automated using an Opentrons OT-2 liquid handler and an Agilent plate reader. This automation would reduce the time for that step from 2 hours to 1 hour for 96 samples and from 2 hours to 1.5 hours for 192 samples. A second improvement is to interleave the Fragmentation and End Prep step and the Adapter Ligation and cleanup step when processing 2 plates. Since Fragmentation and End Prep ends with an almost hour-long thermocycler incubation, this incubation can start for the second plate right when the first plate is finishing and then complete Adapter Ligation and cleanup for the first plate while the second plate is in the thermocycler. Figure 7B shows a diagram of how two plates would be analyzed once these improvements are implemented. The sample to answer time once these modifications are in place is shown in Table 36.

Table 36. Sample to Answer time summary once Quantification, Normalization and Pooling is automated on OT-2 and Fragmentation and End Prep is interleaved with Adapter Ligation and clean up when processing two plates.

Sample to Answer Times	NextSeq 2000 96 (planned)		NextSeq 2000 192 (planned)	
Sequencer Flow Cell	NextSeq 2000 P2 2 x 50 bp		NextSeq 2000 P3 50 bp	
Number of Samples	96		192	
Times in hours	Preparation Time	Sample to Answer Time	Preparation Time	Sample to Answer Time
Nucleic Acid Extraction	0.5	1	0.5	1
Reverse Transcription and Template Switching	0.2	1.9	0.2	2.2
cDNA amplification by PCR	0.3	1.1	0.4	1.1
cDNA cleanup	0.3	1.7	0.4	3.4
Fragmentation and End Prep	0.2	1	0.3	1
Adapter Ligation and clean up	0.3	1.1	0.4	2.2
Library PCR and clean up	0.2	1	0.3	1.1
Quantification, Normalization, Pooling, Cleanup	0.3	1	0.4	1.2
Sequencing	0.5	13	0.5	11
Bioinformatics	0	0.3	0	0.4
Total Sample to Answer Time		23.1		24.6

[00167] Prior to utilizing real samples for validating the automation, some embodiments performed many runs of the automation using water to validate that the system was performing as expected. The automation validation experiment generated a plate 96 remnant samples where data from the Roche ePlex for each sample is available. For these 96 samples, both manual and automated library preparation are performed. Both manually and automated prepared samples are sequenced together. Any positives that are next to other positives and not consistent with the Roche ePlex results are removed. The same analysis is performed for the other pathogens and the automation is consistent with the manual process and reduces the contamination. Figure 8A through 8I show the results from automation validation for three additional pathogens. Each row is a different pathogen: (a)-(c) Human Coronavirus. (d)-(f) Parainfluenza 1. (g)-(i) Human Metapneumovirus. The first column (a), (d), (g) are locations of positives confirmed by Roche ePlex. The second column (b), (e), (h) are locations predicted by the assay using the manual library preparation protocol. The third column are locations predicted by the

assay using the automated library preparation protocol. Blue positions are true positive predictions and red are false positives.

[00168] Many embodiments provide methods to scale to a much larger number of samples. The bottleneck for the number of samples is not the capacity of the sequencer, but the effort to construct a library. This process takes multiple hours and has many steps. Each step has to be performed for each sample. In various methods in accordance with many embodiments, molecular barcodes can be added to each sample at the first step, the samples can be mixed together, and all of the remaining steps can be performed on the mixture. Several embodiments implement this method using the BRB-Seq library. Some embodiments repeat the clinical and analytical validations using this technique. The technique is not as efficient as the standard approach and can require a bit more sequencing than the standard. Using this approach, several embodiments can find viral pathogens in the samples. The advantage of this approach is that the number of samples to be processed can be substantially increased and the time it takes to run the samples can be substantially decreased. Some embodiments can combine about 96 or 384 samples into a single mixture and process them all together. If this approach is combined with the automation approach that can prepare 192 samples described above, some embodiments can then process 192 mixtures of 384 samples or over 70,000 samples in close to 24 hours.

[00169] This mixture of samples has improved performance when depletion approaches are applied in accordance with many embodiments. Specifically, some embodiments show that the mixture of sample approach benefits from CRISPR depletion unlike the standard approach.

[00170] Several embodiments perform the methods on microfluidic liquid handlers using instruments such as: Tecan MagicPrep® and Revvity BioQule®. Using these instruments, there is much less hands on time for an operator to apply the technique. These instruments can process up to 8 samples at a time. If combining the mixture of barcoded samples approach with this approach, certain embodiments can process 384 * 8 or over 3000 samples on a microfluid liquid handler.

[00171] Several embodiments apply the methods to multiple sample types. Some embodiments show that this technique can apply to any clinical sample type by showing

that it applies to a set of representative sample types including (but not limited to) urine, blood, saliva, sputum, nasal swab, bronchoalveolar lavage (BAL), tissue, skin swab. As these sample types are representative of the breadth of sample types, this technique can be applied to all clinical sample types.

[00172] Some embodiments provide that the methods can lead to full reconstruction of the virus genomes using sequencing assembly. Several embodiments demonstrate this by showing that we these methods can reconstruct viruses that were unknown to be present in the sample at the time.

[00173] In various embodiments, the methods can be combined with targeted next generation sequencing diagnostics in the same sequencing run. For example, the Illumina MiniSeq® 5 sample run can also process several hundred targeted diagnostics.

Examples

[00174] Example 1: A method of detecting a pathogen comprising,
obtaining a plurality of nucleic acids from a sample;
sequencing the plurality of nucleic acids to obtain a plurality of reads;
analyzing the plurality of reads using a bioinformatic process; and
identifying the pathogen from the analyzed reads.

[00175] Example 2: The method of example 1, further comprising obtaining the sample using at least one of: a swab, a Q-tip, a wipe, and a cloth.

[00176] Example 3: The method of example 1 or 2, wherein the sample is a tissue, an organ, a bodily fluid, an object, a surface, or a container.

[00177] Example 4: The method of example 1, or 2, or 3, wherein the sample is at least one of: a turbinate swab, a nasal pharyngeal swab, and saliva.

[00178] Example 5: The method of any one of examples 1 to 4, wherein the sample is a clinical sample; wherein the clinical sample is at least one of: urine, a tissue, a skin swab, bronchoalveolar lavage (BAL), sputum, and blood.

[00179] Example 6: The method of any one of examples 1 to 5, wherein the sample is an upper respiratory specimen or a lower respiratory specimen from a subject.

[00180] Example 7: The method of any one of examples 1 to 6, wherein the obtaining step obtains the plurality of nucleic acids using at least one nucleic acid extraction kit.

[00181] Example 8: The method of any one of examples 1 to 7, further comprising adding a known quantity of Bacteriophage MS2 to the sample.

[00182] Example 9: The method of any one of examples 1 to 8, further comprising constructing at least one next generation sequencing library.

[00183] Example 10: The method of any one of examples 1 to 9, wherein the at least one next generation sequencing library is at least one of: a DNA library, and an RNA library.

[00184] Example 11: The method of any one of examples 1 to 10, wherein the sequencing step uses the at least one next generation sequencing library.

[00185] Example 12: The method of any one of examples 1 to 11, wherein the analyzing step comprises matching a plurality of sequencing reads to at least one pathogen sequence in a database.

[00186] Example 13: The method of any one of examples 1 to 12, further comprising generating a plurality of simulated reads using a set of curated pathogen sequences; analyzing the plurality of simulated reads using the database to check if the database contains an error; and eliminating a plurality of false reads in the database by analyzing the plurality of simulated reads.

[00187] Example 14: The method of any one of examples 1 to 13, further comprising using the plurality of reads of the pathogen and a plurality of reads of a control added prior to nucleic acid extraction or a plurality of reads of a control added after nucleic acid extraction to quantify an amount of the sample.

[00188] Example 15: The method of any one of examples 1 to 14, wherein the control is Bacteriophage MS2.

[00189] Example 16: The method of any one of examples 1 to 15, wherein the pathogen is a pathogenic organism, a bacterium, a virus, or a fungus.

[00190] Example 17: The method of any one of examples 1 to 16, further comprising processing a plurality of samples at a time using automation.

[00191] Example 18: The method of any one of examples 1 to 17, further comprising processing a plurality of samples at a time using automation by interleaving a plurality sets of samples to increase a number that is processed on an automated instrument.

[00192] Example 19: The method of any one of examples 1 to 18, further comprising processing a plurality of samples using at least one microfluidic liquid handler.

[00193] Example 20: The method of any one of examples 1 to 19, wherein automation eliminates cross-sample contamination when processing the plurality of samples.

[00194] Example 21: The method of any one of examples 1 to 20, further comprising adding at least one molecular barcode to each of the plurality of samples, and mixing the plurality of samples.

[00195] Example 22: The method of any one of examples 1 to 21, wherein more than 90 samples are processed together at a time.

[00196] Example 23: The method of any one of examples 1 to 22, wherein more than 200 samples are processed together at a time.

[00197] Example 24: The method of any one of examples 1 to 23, wherein more than 1000 samples are processed together at a time.

[00198] Example 25: The method of any one of examples 1 to 24, wherein the preparation for sequencing and analyzing are performed on a microfluidic liquid handler.

[00199] Example 26: The method of any one of examples 1 to 25, further comprising a depletion process configured to improve a performance of the mixture of barcoded plurality of samples.

[00200] Example 27: The method of any one of examples 1 to 26, further comprising combining a plurality of targeted diagnostic samples with a plurality of metagenomic diagnostic samples in a same sequencing process.

[00201] Example 28: The method of any one of examples 1 to 27, further comprising using a nucleic extraction process that is optimized for extracting bacteria and viruses compared to host nucleic acids.

[00202] Example 29: The method of any one of examples 1 to 28, wherein sample to result using short read sequencers is completed in less than 11 hours.

[00203] Example 30: A method of a bioinformatic analysis comprising,

generating a plurality of simulated genomic reads using a plurality of pathogen genomic sequences;

eliminating a plurality of false reads in a database by analyzing the plurality of simulated genomic reads;

comparing a plurality of sequencing reads from a sample with the database; and

identifying a pathogen from the sample; wherein the eliminated false reads in the database improves identification accuracy.

[00204] Example 31: The method of example 30, further comprising analyzing the plurality of simulated genomic reads using the database to check if the database contains an error.

[00205] Example 32: The method of example 30 or 31, further comprising generating a plurality reads of a known quantity of Bacteriophage MS2 and using the plurality reads of the known quantity of Bacteriophage MS2 and the plurality of sequencing reads to quantify an amount of the sample.

[00206] Example 33: The method of example 30, or 31, or 32, further comprising extracting a plurality of nucleic acids from the sample.

[00207] Example 34: The method of any one of examples 30 to 33, wherein the extracting step uses at least one nucleic acid extraction kit.

[00208] Example 35: The method of any one of examples 30 to 34, further comprising using a next generation sequencer to generate the plurality of sequencing reads.

[00209] Example 36: The method of any one of examples 30 to 35, wherein the pathogen is a bacterium or a virus.

[00210] Example 37: The method of any one of examples 30 to 36, wherein the pathogen is reconstructed using a genome assembly technique.

[00211] Example 38: The method of any one of examples 30 to 37, wherein eliminating the plurality of false reads uses a manual process.

[00212] Example 39: The method of any one of examples 30 to 38, wherein eliminating the plurality of false reads uses a computer assisted process.

[00213] Example 40: The method of any one of examples 30 to 39, wherein the computer assisted processes uses artificial intelligence.

[00214] Example 41: The method of any one of examples 30 to 40, further comprising using a control sequence that is an organism other than Bacteriophage MS2.

[00215] Example 42: The method of any one of examples 30 to 41, further comprising adding a control RNA or DNA after nucleic acid extraction to generate a plurality of simulated genomic reads.

[00216] Example 43: The method of any one of examples 30 to 42, wherein the database used has publicly available sequencing data that is curated prior to analysis using efficient algorithms.

DOCTRINE OF EQUIVALENTS

[00217] As can be inferred from the above discussion, the above-mentioned concepts can be implemented in a variety of arrangements in accordance with embodiments of the invention. Accordingly, although the present invention has been described in certain specific aspects, many additional modifications and variations would be apparent to those skilled in the art. It is therefore to be understood that the present invention may be practiced otherwise than specifically described. Thus, embodiments of the present invention should be considered in all respects as illustrative and not restrictive.

[00218] As used herein, the singular terms "a," "an," and "the," may include plural referents unless the context clearly dictates otherwise. Reference to an object in the singular is not intended to mean "one and only one" unless explicitly so stated, but rather "one or more."

[00219] As used herein, the terms "approximately," and "about" are used to describe and account for small variations. When used in conjunction with an event or circumstance, the terms can refer to instances in which the event or circumstance occurs precisely as well as instances in which the event or circumstance occurs to a close approximation. When used in conjunction with a numerical value, the terms can refer to a range of variation of less than or equal to ± 10% of that numerical value, such as less than or equal to ±5%, less than or equal to ±4%, less than or equal to ±3%, less than or equal to ±2%, less than or equal to ±1 %, less than or equal to ±0.5%, less than or equal to ±0.1 %, or less than or equal to ±0.05%.

[00220] Additionally, amounts, ratios, and other numerical values may sometimes be presented herein in a range format. It is to be understood that such range format is used for convenience and brevity and should be understood flexibly to include numerical values explicitly specified as limits of a range, but also to include all individual numerical values or sub-ranges encompassed within that range as if each numerical value and sub-range is explicitly specified. For example, a ratio in the range of about 1 to about 200 should be

understood to include the explicitly recited limits of about 1 and about 200, but also to include individual ratios such as about 2, about 3, and about 4, and sub-ranges such as about 10 to about 50, about 20 to about 100, and so forth.

WHAT IS CLAIMED IS:

1. A method of detecting a pathogen comprising,
   obtaining a plurality of nucleic acids from a sample;
   sequencing the plurality of nucleic acids to obtain a plurality of reads;
   analyzing the plurality of reads using a bioinformatic process; and
   identifying the pathogen from the analyzed reads.

2. The method of claim 1, further comprising obtaining the sample using at least one of: a swab, a Q-tip, a wipe, and a cloth.

3. The method of claim 1, wherein the sample is a tissue, an organ, a bodily fluid, an object, a surface, or a container.

4. The method of claim 1, wherein the sample is at least one of: a turbinate swab, a nasal pharyngeal swab, and saliva.

5. The method of claim 1, wherein the sample is a clinical sample; wherein the clinical sample is at least one of: urine, a tissue, a skin swab, bronchoalveolar lavage (BAL), sputum, and blood.

6. The method of claim 1, wherein the sample is an upper respiratory specimen or a lower respiratory specimen from a subject.

7. The method of claim 1, wherein the obtaining step obtains the plurality of nucleic acids using at least one nucleic acid extraction kit.

8. The method of claim 1, further comprising adding a known quantity of Bacteriophage MS2 to the sample.

9. The method of claim 1, further comprising constructing at least one next generation sequencing library.

10. The method of claim 9, wherein the at least one next generation sequencing library is at least one of: a DNA library, and an RNA library.

11. The method of claim 9, wherein the sequencing step uses the at least one next generation sequencing library.

12. The method of claim 1, wherein the analyzing step comprises matching a plurality of sequencing reads to at least one pathogen sequence in a database.

13. The method of claim 12, further comprising generating a plurality of simulated reads using a set of curated pathogen sequences; analyzing the plurality of simulated reads using the database to check if the database contains an error; and eliminating a plurality of false reads in the database by analyzing the plurality of simulated reads.

14. The method of claim 8, further comprising using the plurality of reads of the pathogen and a plurality of reads of a control added prior to nucleic acid extraction or a plurality of reads of a control added after nucleic acid extraction to quantify an amount of the sample.

15. The method of claim 14, wherein the control is Bacteriophage MS2.

16. The method of claim 1, wherein the pathogen is a pathogenic organism, a bacterium, a virus, or a fungus.

17. The method of claim 1 further comprising processing a plurality of samples at a time using automation.

18. The method of claim 1, further comprising processing a plurality of samples at a time using automation by interleaving a plurality sets of samples to increase a number that is processed on an automated instrument.

19. The method of claim 1, further comprising processing a plurality of samples using at least one microfluidic liquid handler.

20. The method of claim 18, wherein automation eliminates cross-sample contamination when processing the plurality of samples.

21. The method of claim 18, further comprising adding at least one molecular barcode to each of the plurality of samples, and mixing the plurality of samples.

22. The method of claim 18, wherein more than 90 samples are processed together at a time.

23. The method of claim 18, wherein more than 200 samples are processed together at a time.

24. The method of claim 18, wherein more than 1000 samples are processed together at a time.

25. The method of claim 19, wherein the preparation for sequencing and analyzing are performed on a microfluidic liquid handler.

26. The method of claim 21, further comprising a depletion process configured to improve a performance of the mixture of barcoded plurality of samples.

27. The method of claim 1, further comprising combining a plurality of targeted diagnostic samples with a plurality of metagenomic diagnostic samples in a same sequencing process.

28. The method of claim 1, further comprising using a nucleic extraction process that is optimized for extracting bacteria and viruses compared to host nucleic acids.

29. The method of claim 1, wherein sample to result using short read sequencers is completed in less than 11 hours.

30. A method of a bioinformatic analysis comprising,
generating a plurality of simulated genomic reads using a plurality of pathogen genomic sequences;
eliminating a plurality of false reads in a database by analyzing the plurality of simulated genomic reads;
comparing a plurality of sequencing reads from a sample with the database; and
identifying a pathogen from the sample; wherein the eliminated false reads in the database improves identification accuracy.

31. The method of claim 30, further comprising analyzing the plurality of simulated genomic reads using the database to check if the database contains an error.

32. The method of claim 30, further comprising generating a plurality reads of a known quantity of Bacteriophage MS2 and using the plurality reads of the known quantity of Bacteriophage MS2 and the plurality of sequencing reads to quantify an amount of the sample.

33. The method of claim 30, further comprising extracting a plurality of nucleic acids from the sample.

34. The method of claim 33, wherein the extracting step uses at least one nucleic acid extraction kit.

35. The method of claim 30, further comprising using a next generation sequencer to generate the plurality of sequencing reads.

36. The method of claim 30, wherein the pathogen is a bacterium or a virus.

37. The method of claim 30, wherein the pathogen is reconstructed using a genome assembly technique.

38. The method of claim 30, wherein eliminating the plurality of false reads uses a manual process.

39. The method of claim 30, wherein eliminating the plurality of false reads uses a computer assisted process.

40. The method of claim 39, wherein the computer assisted processes uses artificial intelligence.

41. The method of claim 30, further comprising using a control sequence that is an organism other than Bacteriophage MS2.

42. The method of claim 30, further comprising adding a control RNA or DNA after nucleic acid extraction to generate a plurality of simulated genomic reads.

43. The method of claim 30, wherein the database used has publicly available sequencing data that is curated prior to analysis using efficient algorithms.

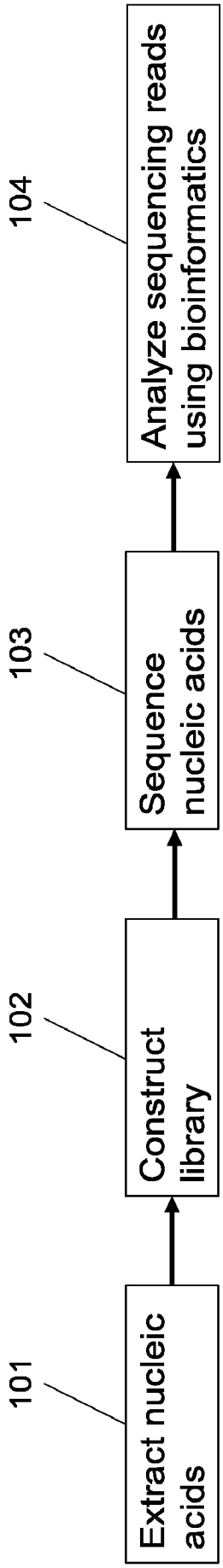

101
Extract nucleic acids
102
Construct library
103
Sequence nucleic acids
104
Analyze sequencing reads using bioinformatics

FIG. 1

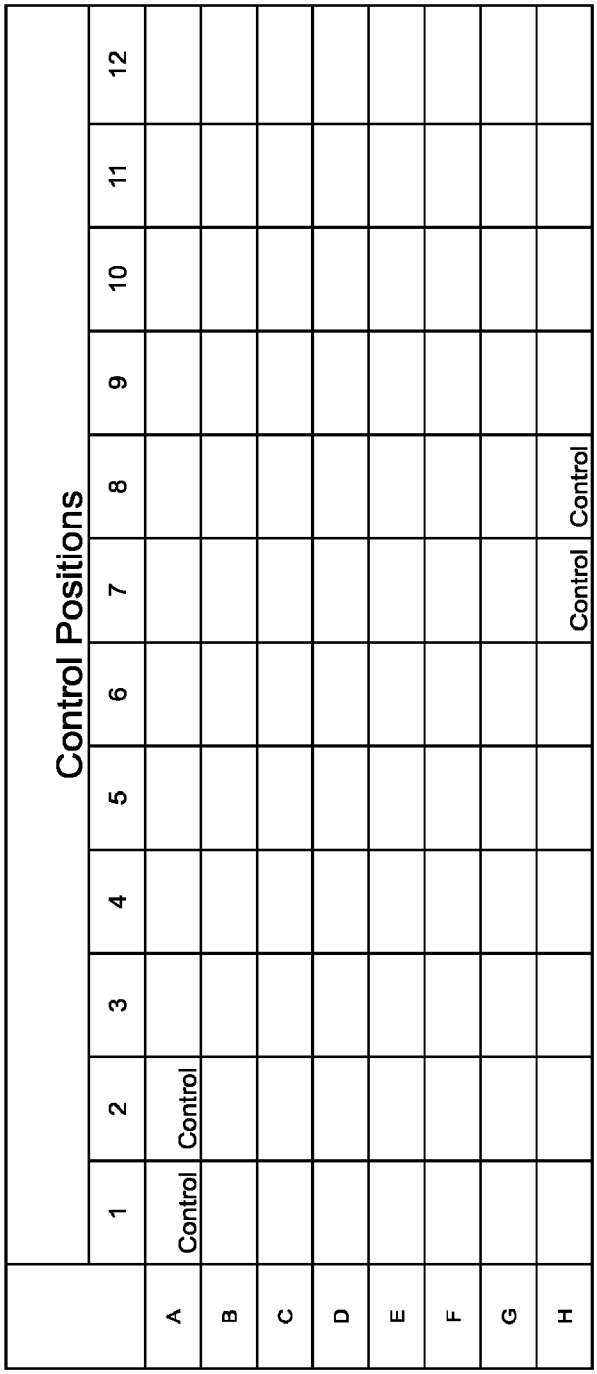

Control Positions
1
2
3
4
5
6
7
8
9
10
11
12
A
B
C
D
E
F
G
H
Control
Control
Control
Control

FIG. 2

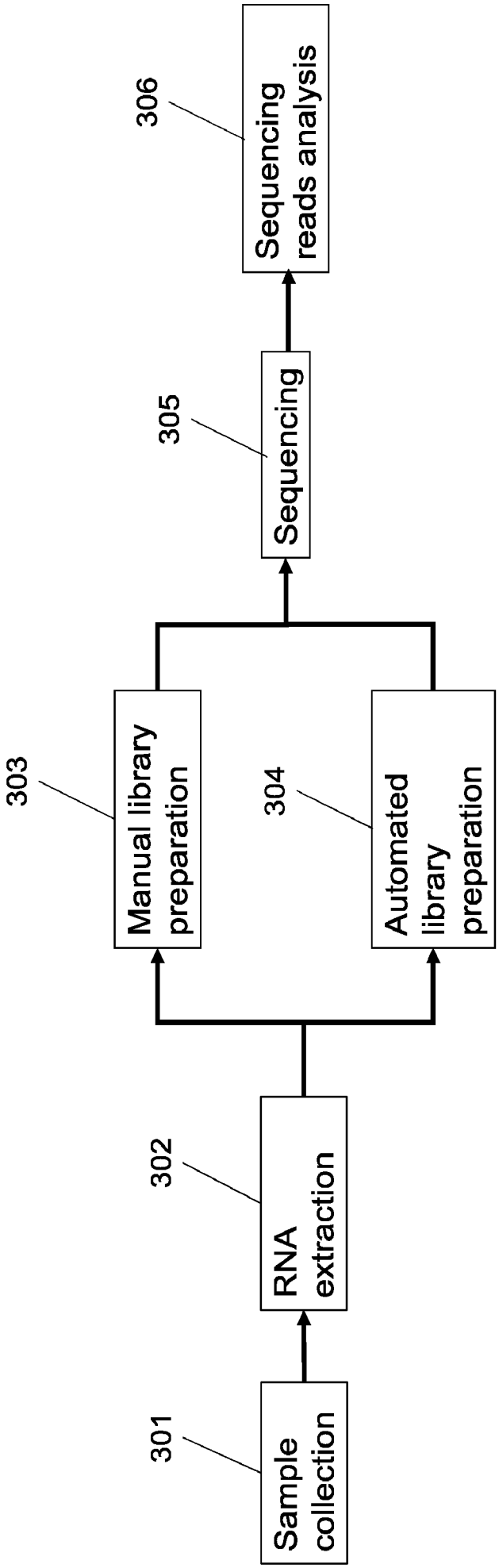

301
Sample collection
302
RNA extraction
303
Manual library preparation
304
Automated library preparation
305
Sequencing
306
Sequencing reads analysis

FIG. 3

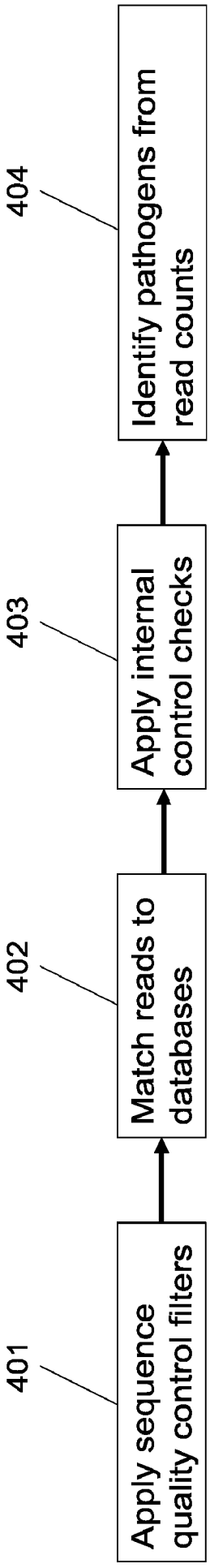

401
Apply sequence quality control filters
402
Match reads to databases
403
Apply internal control checks
404
Identify pathogens from read counts

FIG. 4

Sample	unclassified	classified	human	bacteria	virus	RSV	influenza A	SARS-CoV-2	% SC2 *1k	PCR Result
1	75441	1354221	1308174	41253	2466	0	0	2180	1.60978156	Positive
2	117510	1609527	1588136	13306	4559	0	0	4257	2.64487641	Positive
3	3105117	29414286	28598439	705369	52680	0	0	0	0	Negative
4	74444	858107	816067	27715	291	8	0	0	0	Negative
5	2713365	9528279	5860046	162911	3445895	2126285	0	0	0	Negative
6	121025	1522143	1504595	11699	83	0	0	0	0	Negative
7	149222	1193083	1172412	15388	252	0	0	0	0	Negative
8	166585	1443298	1393233	29286	440	0	0	0	0	Negative
9	128861	1587739	1570126	15233	122	0	0	0	0	Negative
10	268949	3732712	3680275	21579	19834	128	0	0	0	Negative
11	34070	594534	581135	12176	208	0	0	104	0.17492692	Positive
12	1461261	9536174	9362258	115440	1840	0	0	0	0	Negative
13	63064	879271	863253	7650	57	0	0	0	0	Negative
14	12649	99595	97731	1462	14	0	0	0	0	Negative
15	675750	5822757	5724787	84154	577	0	0	0	0	Negative
16	62341	525141	513900	7057	177	1	0	0	0	Negative
17	124811	1238258	1196532	36291	180	0	0	0	0	Negative
18	144165	3283237	3264784	13125	184	0	0	7	0.00213204	Negative
19	72649	1676754	1444734	224826	232	0	0	0	0	Negative
20	132114	1806077	1767374	35179	293	0	0	114	0.06312023	Positive

FIG. 5

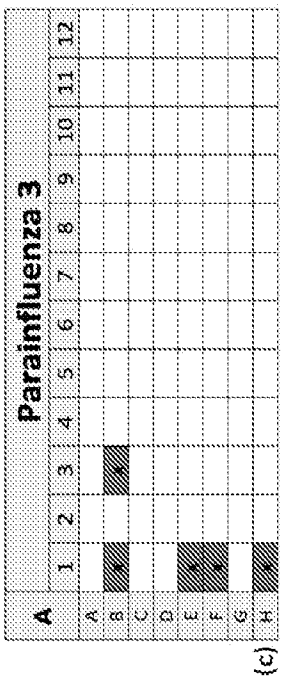


FIG. 6C

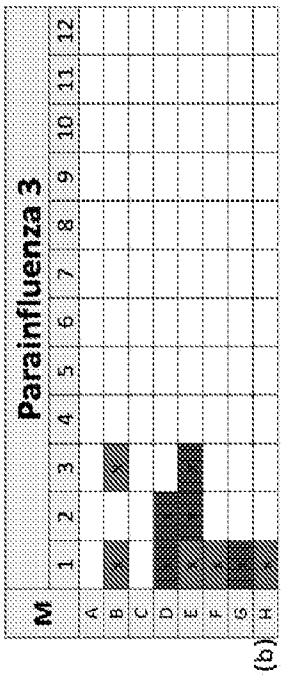


FIG. 6B

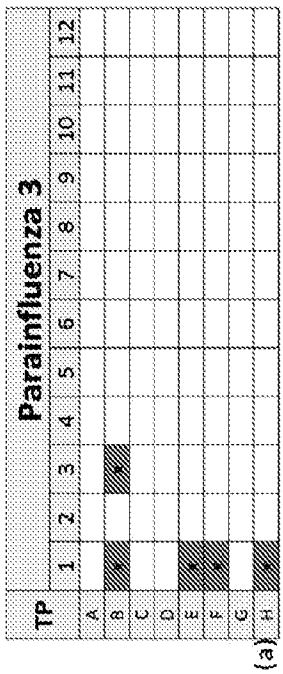


FIG. 6A

FIG. 7A

FIG. 7B

E – Extraction
L1 – Reverse Transcription and Template Switching
L2 – cDNA amplification by PCR
L3 – cDNA cleanup
L4 – Fragmentation and End Prep
L5 – Adapter Ligation and clean up
L6 – Library PCR and clean up
L7 – Quantification, Normalization and Pooling
S – Sequencing

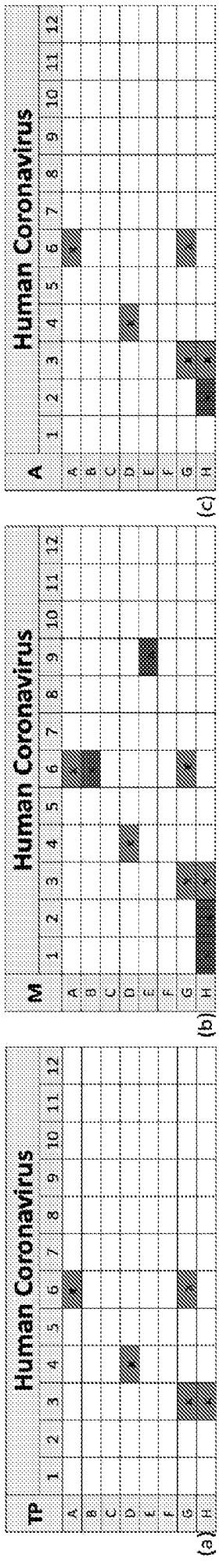
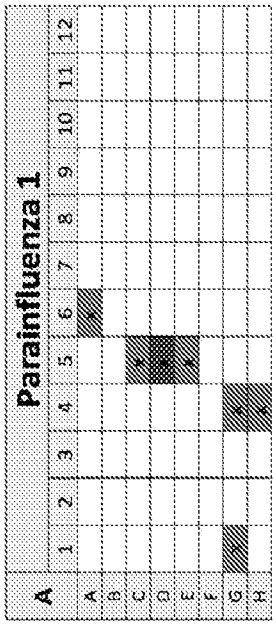


FIG. 8A

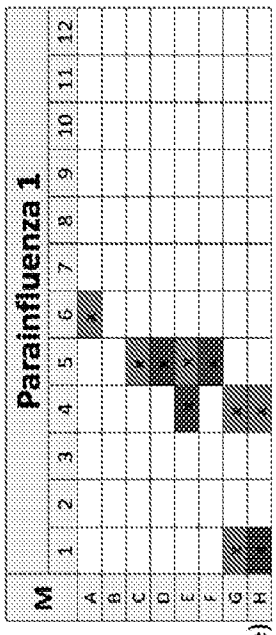
FIG. 8B

FIG. 8C



(d)

FIG. 8D



(e)

FIG. 8E

(f)

FIG. 8F

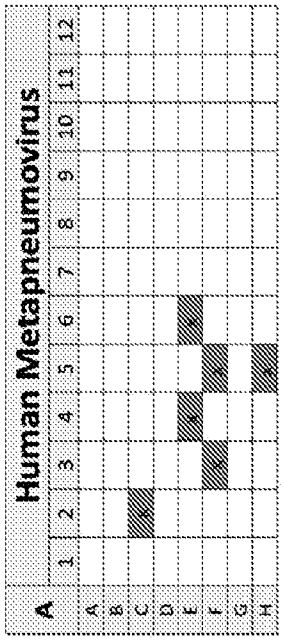


FIG. 8I

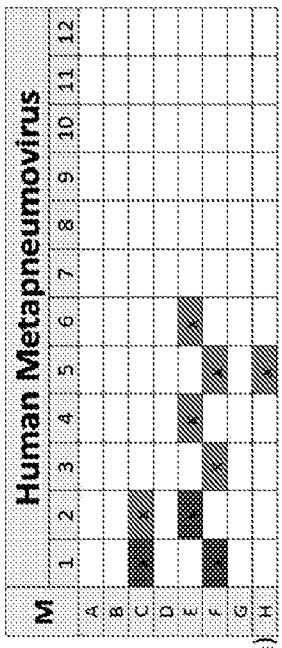


FIG. 8H

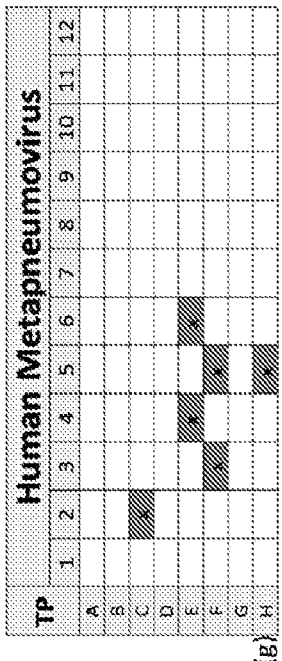


FIG. 8G