



(51) International Patent Classification:

C12N 15/10 (2006.01) G06F 19/24 (2011.01)
G06F 19/14 (2011.01)

94041 (US). **LI, Xiyan**; 218 Murlagan Avenue B, Mountain View, CA 94043 (US). **SNYDER, Michael**; 949 Cottrell Way, Stanford, CA 94305 (US).

(21) International Application Number:

PCT/US2018/028542

(74) **Agent: JACOBS, Ron et al.**; 438 Cambridge Avenue, Suite 225, Palo Alto, CA 94306 (US).

(22) International Filing Date:

20 April 2018 (20.04.2018)

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/488,119 21 April 2017 (21.04.2017) US
62/488,256 21 April 2017 (21.04.2017) US
62/617,471 15 January 2018 (15.01.2018) US

(71) **Applicant: THE BOARD OF THE TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY** [US/US]; 3000 El Camino Real, Building Five, Third Floor, Palo Alto, CA 94306 (US).

(72) **Inventors: JIANG, Chao**; 712 Astor Court, Mountain View, CA 94043 (US). **WANG, Xin**; 3943 El Camino Real#16, Palo Alto, CA 94306 (US). **WANG, Ting**; 10021 Carmona Court, Cupertino, CA 95014 (US). **INLORA, Jingga**; 138 W Dana Street, Apt. A, Mountain View, CA

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) **Title:** ULTRA-SENSITIVE AND UNIVERSAL SPECIES DETECTION PIPELINE FOR NEXT GENERATION SEQUENCING DATA

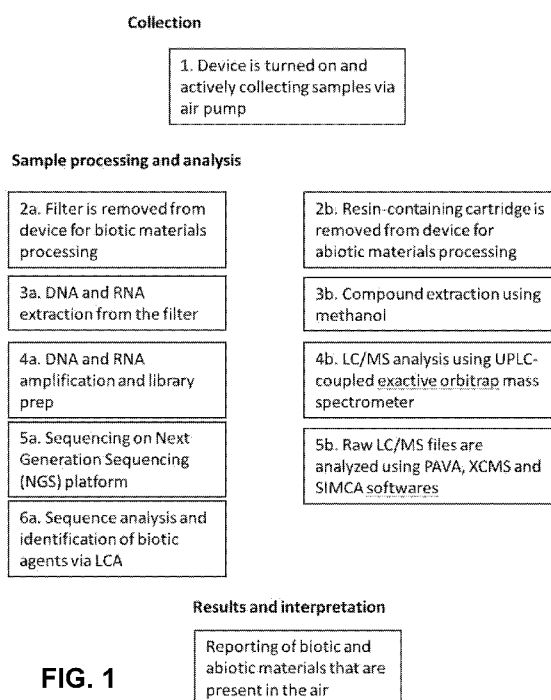


FIG. 1

(57) **Abstract:** A presumption-free pipeline is provided that employs experimental and analytic modules to profile samples, including clinical samples, regardless of the complexity and abundance. The experimental and analytic modules can work independently of each other if the user so desired. The experimental module for ultra-sensitive DNA/RNA extraction and sequencing can be used to extract information from any samples to feed into analytical pipelines chosen by the user. Alternatively, the analytic module for universal species detection can be fed with data generated with other experimental pipelines and different sequencing platforms.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

**ULTRA-SENSITIVE AND UNIVERSAL SPECIES
DETECTION PIPELINE FOR NEXT GENERATION
SEQUENCING DATA**

5 FIELD OF THE INVENTION

This invention relates to systems, methods and analysis pipelines for species detection based on DNA and RNA fragments.

BACKGROUND OF THE INVENTION

10 The advent of the next generation sequencing (NGS) technology has revolutionized all aspects of biological and medical research, leading to the emergence of several previously inconceivable cutting-edge fields and applications. Among which is the integration of microbiomes -- ubiquitous microbial communities that consist of prokaryotic, eukaryotic, and viral
15 organisms invisible to naked eyes. Studies of microbial communities have revealed previously underappreciated stunning diversity and abundance of microbial organisms in all kinds of sampled places, unraveling complicated intricacies between biotic and abiotic factors in macro- and micro- ecosystems, among which the most prominent being the human body. Studies in human
20 microbiome (microbial communities living on human) have revealed its correlations with a growing list of health problems: allergies, asthma, multiple sclerosis diabetes (types 1 and 2), obesity, autism, inflammatory bowel diseases

and even certain types of cancer. These discoveries have led to the emergence of personalized microbiome treatments, aiming to restore or utilize the microbial communities for the benefit of the patients. Despite these remarkable advances, microbiome studies are generally lacking in at least two major aspects:

1. The lack of sensitivity in rigorously analyzing low-abundance samples collected from air, clear liquid, and certain clinical samples. This is due to the ubiquitous DNA/RNA contamination issues during sample preparation protocols, a problem commonly encountered in the field of forensic sciences. A rigorous decontamination protocol and careful experimental designs are required to achieve higher sensitivity in detection and avoid false-positive detection.

2. The ignorance of other types of organisms that co-exist with the bacterial counterparts in the microbial communities. Non-bacterial microbial organisms include fungi, protozoa, plant pollens, virus especially RNA virus, and other traces of small insects and invertebrate parasites etc. It is known that these non-bacterial organisms are critical to human health and environments. For example, fungi spores and plant pollens can cause allergic reactions and asthma. Virus (especially RNA virus), Fungi, Protozoa, insects, and other types of human parasites can strongly impact human health in numerous ways (Zika and malaria are carried by mosquito, for example).

In this invention we aimed to address these two aspects concurrently and

advance the art with an ultra-sensitive pipeline for universal species detection.

SUMMARY OF THE INVENTION

5 Provided is a presumption-free pipeline that employs experimental and analytic modules to profile samples, including clinical samples, regardless of the complexity and abundance (with unparalleled detection sensitivity down to single microbial cell level, equivalent to 1/500 of a typical human cell in size and 1/1000 in nucleic acid content).

10 This invention pertains to computer-implemented software method as a pipeline that includes a fully custom built genomic database and its accompanying taxonomy database. The pipeline uses the known search algorithm BLASTN to search DNA fragments against the fully custom built genomic database, and then uses an implementation we developed of lowest common ancestor (LCA)
15 algorithm and the taxonomy database to classify fragments.

Experimental Module (Sample extraction and sequencing library preparation)

We have developed a streamlined procedure to process any samples for ultra-
20 sensitive sequencing analysis. Starting from any samples that contain the microbial communities of interests, our experimental pipeline can efficiently

break down any bacterial, fungi, plant, and animal cells, even when embedded in other scaffolds such like soil, human feces, and filters.

The pipeline allows concomitant extraction of DNA and RNA from a single
5 sample. All reagents went through a thorough decontamination procedure to ensure minimal foreign contaminating DNA/RNA introduced. Based on the yield of the extraction step, we include an optional amplification step for both DNA and RNA. Specifically, for DNA, we perform isothermal multiple displacement amplification (MDA) adapted from single-cell studies. For RNA,
10 we perform isothermal RNA linear amplification coupled with rRNA depletion. This is vastly more superior to the conventional mRNA fishing approach using the poly-A tail as a bait, as viral RNA (genomic vRNA) would not have those features. Finally, DNA and converted cDNA (from RNA) are subjected to an automatable single-tube protocol for efficient library preparation for the next
15 generation sequencing (NGS) platform, the sequencing results of which are fed into our analytical module.

Analytic Module (Computer-implemented Software Method)

Our analytical module is implemented as a computational pipeline that performs
20 deduplication, quality control, *in silico* decontamination, assembly, and taxonomy classification. The taxonomy classification is achieved by the fully custom-built DARWIN database and the accompanying taxonomy database and our implementation of the lowest common ancestor (LCA) algorithm. The

choice of database alone is the most important step in any taxonomy classification related studies, as it is much harder, if not impossible, to classify species that are simply not included in the database (or worse yet, misclassify them). For these reasons, we survey a broad spectrum of organisms spanning
5 across all domains of life in our DARWIN database. To compensate for the potential long computational time due to the inclusiveness of the database, the analytic module includes three searching algorithms that have different trade-offs between time and sensitivity. In addition, we include a continue option for the CPU-intensive database searching step so the user could choose to resume
10 this process in the events of unexpected interruption. Finally, the CPU intensive database searching step is deployable on Cloud Computing platform such like Google Cloud through virtual system encapsulations (docker images) to help with institutions/individuals who do not have access to the cluster computing engine, where the analytic pipeline was originally developed.

15

Independent capability

It should be noted that our experimental and analytic modules can work independently of each other if the user so desired. The experimental module for ultra-sensitive DNA/RNA extraction and sequencing can be used to extract
20 information from any samples to feed into analytical pipelines chosen by the user. Alternatively, the analytic module for universal species detection can be fed with data generated with other experimental pipelines and different sequencing platforms.

Applications

Our ultra-sensitive and universal species detection pipeline has very broad applications, even well beyond the original intended purpose – to study the human and environmental microbiome. In fact, since we survey all domains of life in our database, this pipeline is viable for analyzing extremely diverse biological samples:

- Eukaryotic (animal/plants/fungi etc.) and/or prokaryotic (bacteria/virus etc.),
- Single organism or a mixture of organisms,
- Extreme low abundance or high abundance,
- Free-living cells or trapped in scaffolds such as soil, feces, and filters etc.

Some outstanding examples are:

1. Diagnosing patients infected with unknown pathogens when conventional diagnosis have failed. In situations where the nature of the pathogen(s) is unknown, which could be bacteria, virus, fungi, protozoa, and parasite, or even as a mixture of several organisms, and the health of the patient is at stake, our pipeline provides a urgent one-stop solution to quickly help the medical practitioners to narrow their search range, if not directly identify the culprit(s).

2. Constructing comprehensive profiles of environmental risk factors in personalized medicine. Human diseases are essentially unfortunate consequences of intertwined genetic and environmental risk factors. By tailoring to individual needs, our pipeline can reveal the full spectrum of biotic environmental risk factors that may incite unwanted responses such like allergy and asthma and provide a complete risk factor map for activity guidance.
3. Implementing in epidemic surveillance on infectious microbial pathogens in public spaces. The majority of epidemic breakouts (or the unlikely events of bio-terrorism attacks) are followed by intensive efforts to identify the pathogens from clinical labs, often in weeks or even months. Instead, we can implement our pipeline along with public sampling stations that provide constant surveillance over key traffic hubs or high population density areas. By analyzing the samples periodically or on demand, we can monitor, track, or even catch the developing epidemic situations, enabling quick epidemic intervention.
4. Improving agriculture yield by enhancing and modifying the host-microbiome mutualistic relationships. Humanity is constantly striving to improve the agricultural yield from our crops and livestock. Similar to the human-microbiome relationship, plants and animals also benefit tremendously from their respective microbial communities. Revealing and utilizing these host-microbiome relationships will undoubtedly increase the yield and disease resistance of crops and livestock. Our

pipeline can provide a broad and unbiased summary on the said complicated host-microbiome relationships.

We attribute the following advantages to this invention:

- 5 1. The ability to extract nucleotide information from very low abundance samples (10^1 bacterial cell level) due to our strict decontamination protocols and unbiased amplification protocols.
2. The ability to classify species spanning all domains of life (broad range detection of highly diverse samples). Previous efforts usually only focus on a
- 10 sub-domain of life, mostly bacteria, virus and maybe some fungi.

We could adapt our experimental pipeline to clinical samples where human tissues are dominant. On the other hand, our database is constantly updated and curated to cover all domains of life heuristically. Finally, a visualization module

15 can be developed for the taxonomy report using open-sourced statistical software R.

In one embodiment, the invention is a pipeline detection with the following steps:

- 20 Deduplication, Quality Control, In silico decontamination, Assembly and Taxonomy classification, all implemented by software on a computer system or

one or more computer processors. The steps can be regarded as computer-implemented steps executable on and by a computer system.

For Deduplication, the input to the pipeline is raw sequencing reads in fastq
5 format from the sequencing platforms. The deduplication action or process
pertains to removing exact paired-duplicated reads from the data. Sequences of
each reads are directly hashed and compared to speed up the process. The output
of the action or process is processed de-duplicated sequencing reads in fastq
format.

10

For Quality control, the input is the processed de-duplicated sequencing reads
in fastq format. The quality control action or process is to use e.g.
software Trim_galore, which will remove any remaining sequencing adapters
and low quality bases from the 5 prime and 3 prime ends. Trimmed reads shorter
15 than 30bp are removed all together. The output of the action or process is de-
duplicated, trimmed high quality reads in fastq format.

For In silico decontamination, the input is de-duplicated, trimmed high quality
reads in fastq format. The in silico decontamination action or process is that the
20 processed reads are mapped to the human reference genome hg19 version by
e.g. the bwa-mem algorithm. Reads mapped to the human reference genome are
removed from the sequencing data. The output of the action or process is de-
duplicated, trimmed, nonhuman reads in fastq format.

For Assembly, the input is de-duplicated, trimmed, nonhuman reads in fastq format. The assembly action or process is that the processed reads are assembled de novo using megahit using the metagenome-sensitive preset. The
5 cut-off for DNA contigs are 300 bp, and 200 bp for RNA contigs. Anything shorter than the cut-off are removed. The output of the action or process is assembled contigs from the input reads.

For Taxonomy classification, the input is assembled contigs. The taxonomy
10 classification action or process is that the assembled contigs are searched against a custom built database that covers all kingdoms of life, using e.g. the BLASTN algorithm. A wrapper was written to introduce the continue option and examine the integrity of the BLASTN results. The BLASTN results are parsed using custom-implemented LCA algorithm to achieve a balance between
15 sensitivity and specificity of classification methods. The results are further parsed through a custom written taxonomy report script which generates taxonomy abundance information at all taxonomy levels, in addition to listing species separately for each kingdom of life. Finally, identity of contigs to reference genomes are retained and display at the species level to facilitate
20 the confidence of taxonomy assignment. The output of the action or process is BLASTN results, LCA results, Taxonomy results.

In one embodiment, the invention is an experimental pipeline for pan-domain species nucleotide extraction and next generation sequencing library preparation. In this pipeline, the following steps are included:

- preparing DNA/RNA genome libraries from collected metagenome samples
5 at a single bioparticle sensitivity,

wherein a radiation-based reagent-decontamination protocol is implemented to remove any traces of amplifiable nucleotides from reagents,

wherein a sterilized hydrophobic or hydrophilic filter membrane is used to
10 collect nano-gram-level amount of airborne biological particulates at the minimum in between 12 to 72 hours,

wherein, with an amount range above 1 ng of nucleotides of airborne biological samples, extracting DNA and RNA concurrently from the collected samples of any nature by a chemistry/physics-based pipeline at a
15 single bacterial particle sensitivity,

wherein a non-PCR-based linear amplification technique is applied for DNA and RNA after converted to complementary DNA, and

wherein, with limited RNAs, depleting most ribosome RNA with selective primers for preferential enrichment of non-ribosome RNA sequences; and
20 • using cocktail enzyme to fragmentize DNA/cDNA generated to minimize the loss of nucleotide materials.

- In another embodiment, the invention is a pan-domain species detection bioinformatics pipeline for next generation sequencing data. In this pipeline, the following steps are included, which are executed by a computer or one or more computer processors:
- 5 • building a non-redundant genome sequence database incorporating publicly available genome sequences and/or select organisms of interest;
 - building a taxonomy database to store taxonomy information for every sequence in the database;
 - creating de-duplicated sequencing reads in fastq format by removing exact
10 paired-duplicated reads from raw sequencing reads;
 - trimming the de-duplicated sequencing reads;
 - mapping the trimmed de-duplicated sequencing reads to a human reference genome therewith creating trimmed de-duplicated non-human reads in fastq format;
 - 15 • creating assembled contigs from the trimmed de-duplicated non-human reads in fastq format by a rigorous reads assembly method with specific parameters to minimize chimeric assembly;
 - classifying the assembled contigs by searching against the non-redundant genome sequence database; and
 - 20 • outputting the classified results.

BRIEF DESCRIPTION OF THE DRAWINGS

- FIG. 1** shows according to an exemplary embodiment of the invention an overall flowchart of detection of air biotic and abiotic particulates.
- FIG. 2** shows according to an exemplary embodiment of the invention how UV treatment eliminates background contaminating DNA in NTC. The amplification curve of the 10^3 e. coli cells remains unchanged.
- FIG. 3** shows according to an exemplary embodiment of the invention that the experimental pipeline can extract sufficient information from down to 10 (or even less) bacterial cells. Left, different amount of e. coli cells were spiked in before extraction and their numbers are evaluated by plating them on LB plates and counting. Right, the approximate amount of DNA and RNA from a typical E. coli cell as compared to a typical mammalian cell (HeLa cell).
- FIGs. 4A-C** show according to an exemplary embodiment of the invention spike-in Phi174 provides a direct method to evaluate the original amount of sample in DNA (**FIGs. 4A-B**) and RNA (**FIG. 4C**).
- FIG. 5** shows according to an exemplary embodiment of the invention minimal amount of reads map to Phi174 in actual samples.
- FIG. 6** shows according to an exemplary embodiment of the invention the detection pipeline as being highly reproducible between filters collected at the same time and same location. The bottom figure

shows a similar representation of the two figures about on the y-axis the abundance and the x-axis the individual genus (similar organism).

FIG. 7 shows according to an exemplary embodiment of the invention an overview of the analytical pipeline.

FIG. 8 shows according to an exemplary embodiment of the invention the Darwin database.

FIG. 9 shows according to an exemplary embodiment of the invention UFC providing complete coverage of organisms that are not covered by conventional microbiome package (to the extent of human knowledge).

FIG. 10 shows according to an exemplary embodiment of the invention that the pipeline has a sensitivity down to 10^2 copies of individual viral species in a mixture of various RNA viruses and bacterial and yeast organisms. (the size of viral genome is 1/1000 E. coli cell).

FIG. 11 shows according to an exemplary embodiment of the invention analysis of 100+ samples collected from a study showing dynamic species abundance covering all domains of life. Listed on the right are some of the potential pathogenic species.

DETAILED DESCRIPTION

Definitions

Next-generation sequencing (NGS), also known as high-throughput sequencing, is a catch-all term used to describe a number of different modern sequencing technologies including:

- Illumina (Solexa) sequencing
- 5 • Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

These technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, which is the main
10 reason we are calling it "next generation sequencing". The massively parallel sequencing technology known as next-generation sequencing (NGS) has revolutionized the biological sciences. With its ultra-high throughput, scalability, and speed, NGS enables researchers to perform a wide variety of applications and study biological systems at a level never before possible.

15 Ultra-sensitive is a term relevant to our experimental part of the invention, where we show that the pipeline is able to extract sufficient information from 10 bacterial cells and 200 viral particles.

Universal is a term relevant to our custom-built databases, which aim to characterize species from all kingdoms of life, including, but not limited to,
20 bacteria, fungi, viruses, plants, animals, archaea, etc.

Overview

The experimental pipeline of this invention is unique in that it is adapted to single-cell level amount of nucleic acids materials from a mixture of diverse organisms. It is noted that it also works if more materials are provided. The details of the steps of pipeline are provided in the **Experimental Protocols** section. Traditionally, single-cell experiments are only carried out in mammalian cells or bacterial cells where single or a few cells of the same species are processed at a time. The experimental pipeline aims to process a diverse mixture of organisms presented in a very small amount of materials (equivalent or less than 1000 microbial cells, which is about single mammalian cell level for materials). This seemingly contradictory situation requires novel experimental and analytical techniques to faithfully deconvolve the population structure. Therefore, preserving the signatures from diverse organisms and reducing the impact of contamination from either human or reagent sources becomes a paramount task to accomplish.

Decontamination methods

To this end, we employ rigorous reagent selections and specific in-lab decontamination protocols. Specifically, we have tested majority of commercially available microbiome extraction kits and adopted the one that has the following two traits: 1, efficiently breaking diverse organisms' cells and

releasing the nucleic acid contents, 2. high reproducibility and minimal material loss when only a small number of cells are provided (According to the supplier, the use of their kit for such a small number of cells were never done and they consider it impossible). Upon receiving the extraction materials, we then aliquot

5 all reagents that do not contain enzymes into 1.5 ml plastic tubes and place them around 3 cm from the 254nm UV radiation source inside a commercial Stratalinker 2400 UV CROSSLINKER for 30 minutes (4000 mwatts/cm²). The amount of UV energy exposed is at least twice as much as required to break at least 99.9% contaminating nucleic acids in the reagents to sub-73bp (Plos

10 ONE), which should have minimal impact on the downstream amplification and library preparation steps. In addition, all personnel are required wear long-sleeve lab coats and face masks, working in a physically separated and designated clean hood when performing the extraction process to minimize human contamination. In a possible variation, the exact amount of UV light

15 exposure and the volume of each aliquot can be adjusted for larger scale of operations.

The successful outcome of decontamination is reflected in the qPCR quantitation results (**FIGs. 2A-B**). In **FIG. 2A**, DNA extraction was performed

20 without the strict decontamination protocol. We found that DNA extracted from 10^3 *E. coli* cells yield virtually no difference in amplification curve when compared to No Template Control (NTC, i.e., DNase-free water). This suggests

that the *E. coli* DNA signals are completely masked by the inherent contaminating DNA (bacterial origin) in the reagents. In comparison, when strict decontamination protocol is implemented, while the amplification curve of 10^3 *E. coli* cells remain mostly unchanged, the amount of DNA in NTC is no longer detectable (cycle $> 34\sim 35$ is considered sub-single molecule level). These results strongly underscores the importance of our strict decontamination protocol prior to handling materials with extremely low amounts of nucleic acids.

10 Amplification methods

The amount of DNA and RNA extracted from our samples are usually in such low quantities that machines such as NanoDrop and Qubit are unable to measure. Thus, the second technological hurdle to overcome is to amplify the nucleic acids to a level where sequencing libraries can be prepared. Commercially available next generation sequencing (NGS) library preparation kits require minimum 1 ng input, which is approximately 1000x more than the amount we obtain from extraction. To this end, we utilize a single-cell Multiple Displacement Amplification kit to amplify DNA. For RNA, a single primer isothermal amplification kit specifically designed to amplify all non-rRNA is used. As most RNA amplification kits are tailored to mRNA, which selectively enrich for RNA that contain poly-A tail, they are unsuitable for our case. This

is because almost all bacterial and viral RNA do not have poly-A tail and therefore will not be amplified. Thus, selecting the broad amplification of all non-rRNA technique is important and will preserve the complex community structures of our samples. Following amplification, DNA and cDNA are
5 converted into sequencing libraries using commercially available kit for next generation sequencing (NGS).

Sensitivity of detection methods

To test the sensitivity of our pipeline, we titrated *E. coli* culture down to 1000,
10 100, and 10 cells and extracted these samples using our pipeline, along with a blank control to monitor the contamination background. Our results show that our pipeline can accurately detect at least down to 10 *E. coli* cells from the sample (**FIG. 3**). This is surprising because conventional extraction methods are catered to extracting bacterial DNA from at least millions of cells. It is most
15 certainly unexpected that our pipeline can also work with minimal input cells and provide good and reliable results (based on communication with senior scientists at company that manufactures the extraction kit).

Spiked-in Evaluation of the amount of materials collected

We also precisely evaluated the actual amount of nucleic acids content in situations where extremely small amount of sample are collected (samples collected from a personal device as disclosed in US Provisional Applications 62/488256 filed on April 21, 2017 and 62/617471 filed on January 15, 2018).

5 To gather samples for this part, we used commercialized RTI device which was intended to collect pollutants on a filter through active sampling from air and measure them using mass-spectrometry. Adapting this strategy, we instead extract biological contents from the filters using our pipeline. To our knowledge, there are no direct methods to reliably measure nucleic acids
10 amount at sub-pg ($<10^{-12}$ g) levels, thus we resort to amplification and sequencing. Prior to DNA amplification, a known amount of *E. coli* phage PhiX174 (5 pg, 500 fg, and 50 fg) is spiked into our sample (in triplicates). The spike-in serve as "ballpark estimates" of the amount of materials initially present. Since our protocol uses random amplification, it is reasonable to
15 assume the final amount ratio between our sample and PhiX174 reflects the actual amount collected. Post-sequencing, the sequencing reads are mapped to human and PhiX174 genomes. Sequencing reads that are non-human and non-PhiX174 are labeled as "others". The number of reads in each category are represented as a percentage of the total reads (**FIGs. 4A-B**). As shown, there is
20 considerable contamination from human resources, as always expected. In ideal cases, if the initial amount of our sample is the same as the spiked-in PhiX174 DNA, the ratio between reads mapped to "others" and PhiX174 should be 1:1. This "others"-to-PhiX174 reads ratio is expected to change proportional to the

initial DNA amount in our sample versus spiked-in PhiX174. At 50 fg of PhiX174 DNA, we noticed that "others"-to-PhiX174 reads ratio are approximately 2:1 for both sample 1 and 2, suggesting that our sample contains more than 50 fg of DNA initially. Increasing PhiX174 (blue) spike-in amount to 500 fg (or 5×10^{-13} g) is sufficient reduce the percentage read of "others" to almost 0. These results suggest that our original extracted amount (samples 1 and 2) are between the range of 50 fg – 500 fg, equivalent to 10 – 100 *E. coli* cells. This assessment is in agreement with our sensitivity evaluation from **FIG. 1** and **FIG. 2**. Similar conclusions can be drawn from the RNA spike-in samples and tests as well (**FIG. 4C**). It is known RNA is less stable than DNA when exposed to environment, which may underestimate the actual number of cells collected. Finally, it is worth noting that we rarely detect PhiX174 in actual samples collected, as shown for both DNA and RNA in **FIG. 5**, and hence was chosen as the spike-in material in our experiments.

15

Detection is highly reproducible

Last but not the least, with our rigorous optimizations, our pipeline is highly reproducible. This is demonstrated by our results where the extraction and processing of two air samples collected side by side show up to 0.9 correlation coefficient at the species level (**FIG. 6**).

20

Analytical pipeline descriptions and supporting analyses

The analytical pipeline, or Universal Fragment Classification (UFC) pipeline, is a collection of scripts written in shell and python (**FIG. 7**). For certain tools, such as BLASTN, a wrapper is written to add a few functionalities into the BLASTN package which are essential for the scope of the analysis. This includes automatically splitting input for massively paralleled database searching and associated output parsing, automatic evaluation of the input and output to see if the results are complete, and a resume option in cases where prior searching step were interrupted due to unforeseen circumstances. Searching tens or hundreds of thousands of contigs of nucleic acids through a massive database (the DARWIN database) is a very time consuming and computing-extensive job, therefore careful optimizations need to be done to ensure the running time for each sample is within reasonable time-frame. The detailed descriptions of each step are described below.

15

The detailed steps of the pipeline

DEDUPLICATION - Amplified DNA or RNA samples frequently suffer from data quality issues where abnormally high coverage of certain regions of genome/transcriptome are observed. This is due to the technical nature of amplification techniques. Conventional approaches attempt to first map reads to reference genomes and use the mapping coordinates to determine if they are

20

duplicates. While memory-efficient, this approach is impossible for most microbiome research because such reference genomes simply do not exist. Therefore, an implementation of reference-free deduplication method is introduced in this pipeline. A possible variation is that the program can be
5 rewritten in C++ for extremely large input size.

TRIMMING AND QUALITY CONTROL - This step is carried out using Trim_galore wrapper, which essentially combines the adapter removal tool Cutadapt and NGS quality control tool Fastqc.

10

DEHUMANIZATION - This step is performed using publicly available BWA-mem algorithm to map all reads to the human reference genome. The purpose is to remove the human reads portion (which is always present when samples need to be amplified before library prep, possibly from the sample handler) from
15 the total reads so that the following assembly step is more efficient. A possible variation is that different version of human reference genome could be used and may yield slightly different results.

DE NOVO ASSEMBLY - This step can be executed either by Megahit or SPAdes,
20 both of which are popular *de novo de bruijn* graph assembler for short read NGS

sequencing reads. The purpose of this step is to assemble millions or more reads into separate information-dense “contigs”, similar to piecing jigsaw puzzles together into bigger clusters. This is an essential step in this pipeline because of its role in data reduction and information retention, thereby increasing
5 confidence in the subsequent taxonomy assignment (longer sequence = better confidence in assignment). A possible variation is that the choice of assembly algorithms and parameters are subject to change depending on the length of reads.

10 **SEARCHING AGAINST THE DARWIN DATABASE** - This step is carried out using a BLASTN wrapper, which takes NCBI BLAST as its core and adding functionalities that are essential to the pipeline. The BLAST algorithm is selected for this purpose because it remains the most sensitive algorithm to identify a given DNA/RNA sequence. Different BLAST algorithms can be
15 specified by user depending on the size of input or sensitivity requirements of the analysis.

The choice of database(s) is the most crucial component when it comes to nucleic acids detection and classification. This is because alignment or mapping
20 algorithms use these so-called reference sequences to identify reads or fragments. A poorly chosen database always leads to under-classification and

sometimes even false-classification. Unless the sequences are very similar, it is fundamentally impossible to identify a group of species that are not included in the database (for example, a bacteria database can hardly detect any fungi). Thus, for accurate identification of organisms, a broad database encompassing

5 all domains of life is essential. In addition, the database needs to be carefully curated. Unfortunately, public databases are often non-curated, which often translate into redundancy, low-quality, and sometimes contaminating data (especially in cases where one species live within another). We've addressed these issues by creating the DARWIN database. This database is an extensively

10 expanded version of the NCBI BLAST NT database, which is hosted by the national center for biotechnology information (NCBI) containing nucleic acid information that represents all domains of life. However, unlike NCBI BLAST NT, which focuses more on the broad human health related organisms, DARWIN was created to better represent all domains of life (**FIG. 8**). This

15 includes public and semi-public sources including the NCBI RefSeq database, Genebank database, and JGI portal expert version. Moreover, the data represented in DARWIN are carefully curated. The most current version of this curated database includes all domains of life known to humans, which are broadly divided into the following categories: plants, protozoa, invertebrates,

20 bacteria, archaea, fungi, virus, non-mammal vertebrates and some selected animals (82 million entries, 40,000 species). A possible variation for the DARWIN database, while described here in a holistic way, can be separated

into several domains of life based on which searching operations can be performed with specific focus and much faster computing time.

TAXONOMY ANALYSIS WITH LCA METHOD - The BLAST results from
5 previous steps only provide an overview of what sequences may be, as provided
in a list of potential organisms ranked by a statistical measure called e-value.
However, consideration has been given to this process and simply picking the
one with best e-value is not robust enough. Instead, a phylogeny inspired
algorithm called Lowest Common Ancestor (LCA) algorithm is preferred. In
10 our analytical pipeline this algorithm is implemented along with special
considerations to certain domains of life that do not conform to usual taxonomy
database structures. Accompanying the DARWIN database, a DARWIN
taxonomy database specific to the DARWIN database (and beyond) is also
constructed. The goal of taxonomy database is to provide a unique taxonomy
15 label to each entry in the DARWIN database, which enables fast and accurate
evaluation of taxonomy in the LCA step. In practice, a noticeable amount of
contigs can be unexpectedly assigned to species belonging to different domains
of life at the same time, hinting a possible contamination source in even well
curated databases. This conflict of assignment can be easily ignored if the
20 database does not contain species from different domains of life. A possible
variation can be that the exact rule of assignment is modified depending on
further optimizations.

TAXONOMY REPORT AND ABUNDANCE ESTIMATION - The inferred taxonomy results from the LCA step is compiled and displayed in human readable format. Specifically, the report follows the hierarchical taxonomy rank conventions of NCBI and display the sequencing abundance of each taxonomy rank in aggregate. Abundance estimation is handled in two approaches, median copy number of contigs assigned to each species and aggregate sequencing amount, which reflect different focuses of the analysis. The final report also includes a special section where species belonging to different domains of life are listed separately so one can quickly inspect domains of their interests. A possible variation is a graphic module where results from this step are made into standardized figures can be introduced.

UFC pipeline detects significantly more species than conventional methods

Side-by-side comparison shows that the performance of our analytical pipeline can identify far more (53% against 7% in the example provided, but in cases where samples are dominated by plants, the percentage can be as drastic as 95% against 3%) portions of sequencing information than conventional packages (FCP package is compared here, **FIG. 9**). For RNA sequencing results the UFC can even identify up to 99% of sequencing information.

UFC pipeline detects most spiked-in viral species in a complex mock mixture sample

Furthermore, in a mock community where we mix a panel of 12 different
5 pathogenic viruses with bacteria and yeasts, we can reliably detect almost all
viruses in the mixture despite their genome size being extremely small
compared to bacteria and yeast (**FIG. 10**). In fact, we were able to detect viruses
down to 10^2 copies, equivalent to only 1/10 of a *E. coli* cell, in a mixture of
organisms, simulating a likely real-life scenario.

10

UFC pipeline detects opportunistic pathogens and human related pathogens in real samples

The final demonstration of the process of the pipeline is reflected by the analysis
on more than 100 actual samples as a part of an academic study, where different
15 species covering all domains of life can be detected with dynamic abundance.
Several opportunistic pathogens and even a parasite in one case can be detected
from the samples (**FIG. 11**).

Experimental protocols

Simultaneous biotics DNA and RNA extraction

Filters captured the biotics samples were used for simultaneous DNA and RNA extraction by combination and modification of MO Bio PowerWater DNA and PowerWater RNA extraction kit. We altered the original protocols to allow

5 extraction of DNA and RNA from the same sample.

Detailed extraction protocol is as follows:

1. Before start extraction, use RNA-free reagent to clean the processing surface in the hood.
- 10 2. Warm PWR1 at 55 degrees Celsius for 10 min. Aliquot 990 ul PWR1 (mix well to make sure have the precipitate in the tube) to 1.5 ml tubes, UV-treat for 1800 sec (30 min).
3. Select 7 sample filters and 1 blank filter as negative control. Insert the filter into the 5 ml PowerWater® Bead Tube.
- 15 4. Add 10 microliters β -mercaptoethanol (β ME) to PWR1. Add 1 ml of Solution PWR1 containing β ME to the PowerWater® Bead Tube.
5. Make sure cap is securely tightened on PowerWater® Bead Tube. Secure the PowerWater® Bead Tube horizontally to a MO BIO Vortex Adapter. The tube caps should be oriented pointing toward the center of
- 20 the Vortex Adapter.

6. Vortex at maximum speed for 5 minutes.
7. Centrifuge the tubes $\leq 4000 \times g$ for 1 minute at room temperature.
8. Transfer all the supernatant to a clean 2 ml Collection Tube. Draw up
the supernatant using a 200 microliters pipette tip by placing it down
5 into the beads.
9. Centrifuge at $13,000 \times g$ for 1 minute.
10. Avoiding the pellet, transfer the supernatant to a clean 2 ml Collection
Tube.
11. Add 200 microliters of Solution PWR2 and vortex briefly to mix.
- 10 Incubate at 4 degrees Celsius for 5 minutes.
12. Centrifuge the tubes at $13,000 \times g$ for 1 minute.
13. Avoiding the pellet, transfer the supernatant to two clean 2 ml Collection
Tube (For DNA and RNA extraction). Expect to recover 800-850
microliters of supernatant (400-425 microliters each for DNA or RNA
15 extraction).

For DNA extraction please follow Step 14 to 23. For RNA extraction please
follow Step 24 to 39.

DNA EXTRACTION STEPS:

- 20 14. Add 325 microliters of Solution PW3 and inverse-mix by hand (no
vortex) briefly.

15. Load 500 microliters of supernatant onto a DNA Spin Filter and centrifuge at 13,000 x g for 1 minute. Discard the flow through and repeat until all the supernatant has been loaded onto the DNA Spin Filter.
- 5 16. Place the DNA Spin Filter basket into a clean 2 ml Collection Tube.
17. Shake to mix Solution PW4 before use. Add 600 microliters of Solution PW4 and centrifuge at 13,000 x g for 1 minute.
18. Discard the flow through and add 600 microliters of Solution PW5 and centrifuge at 13,000 x g for 1 minute.
- 10 19. Discard the flow through and centrifuge again at 13,000 x g for 2 minutes to remove residual wash.
20. Place the DNA Spin Filter basket into a clean 2 ml Collection Tube.
21. Add 52 microliters of Solution PW6 to the center of the white filter membrane. Make sure the drop of solution is thoroughly absorbed into
15 the filter membrane. Incubate for 2 minutes at room temperature.
22. Centrifuge at 13,000 x g for 1 minute.
23. Discard the DNA Spin Filter basket. Aliquot DNA in 3 DNA/RNA LoBind tube (15 microliters each) to avoid freeze-thaw cycle. Seal the tube with parafilm. The DNA is now ready for downstream application.

RNA EXTRACTION STEPS:

24. Add 325 microliters of Solution PWR3 and 325 μ l of Solution PWR4.
Then inverse mix by hand (no vortex) briefly. PWR3 and PWR4 is
5 binding buffer optimized for RNA.
25. Load 600 microliters of supernatant onto a RNA Spin Filter and
centrifuge at 13,000 x g for 1 minute. Discard the flow through and
repeat until all the supernatant has been loaded onto the RNA Spin
Filter.
- 10 26. Shake to mix Solution PWR5. Add 600 microliters of Solution PWR5
and centrifuge at 13,000 x g for 1 minute. Discard the flow through.
27. Centrifuge again at 13,000 x g for 1 minute to remove residual wash.
28. Place the Spin Filter basket into a clean 2 ml Collection Tube.
29. Prepare the DNase I Solution, by thawing the volume of DNase I stock
15 enzyme needed according to the number of samples. Per prep, combine
5 microliters of DNase I stock enzyme with 45 microliters of Solution
PWR6. For 8 samples, add 405 microliters PWR6 to 45 microliters of
DNase I stock enzyme (45 microliters aliquot).
30. To the center of the RNA Spin Filter, add 50 microliters of DNase I
20 Solution (prepared by mixing 45 microliters of Solution PWR6 and 5
microliters DNase I stock solution).
31. Incubate at room temperature for 15 minutes.

32. Add 400 microliters Solution PWR7 and centrifuge the column at 13,000 x g for 1 minute.
33. Discard the flow through and add 600 microliters of Solution PWR5 and centrifuge at 13,000 x g for 1 minute.
- 5 34. Discard the flow through and add 600 microliters of Solution PWR4 and centrifuge at 13,000 x g for 1 minute.
35. Discard the flow through and centrifuge again at 13,000 x g for 2 minutes to remove residual wash.
36. Place the RNA Spin Filter basket into a clean 2 ml Collection Tube.
- 10 37. Add 52 microliters of Solution PWR8 to the center of the white filter membrane. Incubate for 2 minutes at room temperature.
38. Centrifuge at 13,000 x g for 1 minute.
39. Discard the RNA Spin Filter basket. Aliquot RNA in 3 DNA/RNA LoBind tube (15 microliters each) to avoid freeze-thaw cycle. Seal the
- 15 tube with parafilm. The RNA is now ready for downstream application.

DNA and RNA amplification

Biotics DNA samples are linearly amplified by the QIAGEN REPLI-g single cell MDA amplification kit with modifications.

- 20 1. 5 microliters of DNA sample was used for the amplification.

2. Add 5 microliters buffer D1 to the DNA sample. Mix by gently vertexing and centrifuge briefly. Incubate at room temperature for 3 mins.
 3. Immediately add 10 microliters stop buffer N1. Mix by gently vertexing and centrifuge briefly.
 4. For each reaction of 20 microliters denatured DNA, add 29 microliters of REPLI-g sc Reaction Buffer and 2 microliters REPLI-g sc polymerase.
 5. Incubate at 30 degrees Celsius for 8h.
 6. Amplified product is cleaned using 1.8x volumes of the Agencourt XP beads to remove reagents from amplification step. Products are washed twice with 80% Ethanol. The final product is ready for library preparation.
- 15 Biotics RNA samples are linearly amplified by NuGEN Technologies, Inc. Ovation RNA-seq system V2 with modifications.

Step-1: First strand cDNA synthesis

1. Add 2 microliters of A1 (First Strand Primer Mix) to a 0.2 ml PCR tube.

2. Add 5 microliters of RNA sample to the primer.
3. Place the tubes in a pre-warmed thermal cycler at 65 °C for 2 min and hold at 4 degrees Celsius.
4. Prepare a master mix by combining 2.5 microliters of A2 (First
5 Strand Buffer Mix) and 0.5 microliters of A3 (First Strand Enzyme Mix) in a 0.5 ml capped tube.
5. Add 3 microliters of the First Strand Master Mix to each tube.
6. Mix by pipetting 5 times, spin and place on ice.
7. Place the tubes in a pre-cooled thermal cycler programmed as
10 follow: 4 degrees Celsius -1 min, 25 degrees Celsius -10 min, 42 degrees Celsius -10 min, 70 degrees Celsius -15 min, hold at 4 degrees Celsius.
8. Remove the tubes and spin to collect condensation and place on ice.

15 Step-2: Second strand cDNA Synthesis

1. Make a master mix by combining 9.7 microliters of B1 (Second
Strand Buffer Mix) and 0.3 microliters of B2 (Second Strand Enzyme Mix)
in a 0.5 ml capped tube.
2. Add 10 microliters of the Second Strand Master Mix to each
20 First Strand reaction tube.
3. Mix by pipetting 5 times, spin and place on ice.

4. Place the tubes in a pre-cooled thermal cycler programmed as follow:

4 degrees Celsius -1 min, 25 degrees Celsius -10 min, 50 degrees Celsius -30 min, 80 degrees Celsius -20 min, hold at 4 degrees Celsius.

5. Remove the tubes and spin to collect condensation and place on ice.

Step-3: Double-strand cDNA was purified with 1.4 volumes of the Agencourt RNAClean XP beads.

Step-4: Purified cDNA were amplified with Single Primer Isothermal Amplification (SPIA).

1. Make a master mix by sequentially combining 20 microliters of C2 (SPIA Buffer Mix), 10 microliters of C1 (SPIA Primer Mix), 10 microliters of C3 (SPIA Enzyme Mix) in a capped tube.

2. Add 40 microliters of the SPIA Master Mix to each tube containing the double-strand cDNA bound to the dried beads. Use a pipette set to 30 microliters and mix thoroughly at least 8-10 times.

3. Place the tubes in a pre-cooled thermal cycler programmed as follow: 4 degrees Celsius -1 min, 47 degrees Celsius -120 min, 80 degrees Celsius -20 min, hold at 4 degrees Celsius.

4. Remove the tube from thermal cycler, spin to collect condensation and place on ice.
5. Transfer the tubes to the magnet and let stand for 5 min to completely clear the solution of beads.
- 5 6. Carefully transfer 40 microliters of the cleared supernatant containing the SPIA cDNA to a fresh tube.

Step-5: SPIA amplified cDNA were amplified with 0.8 volumes of AMPure XP beads.

10 Possible Variations

1. Step-1, more or less RNA samples could be used for the amplification step.
2. Step-3, the volumes of the Agencourt RNAClean XP beads can vary for different downstream applications.
- 15 3. Step-4, SPIA amplification step can vary for 60, 120, 180 min with different applications.
4. Step-5, the volumes of AMPure XP beads can vary for different downstream applications.

NGS library preparation - DNA

DNA library was conducted with KAPA HyperPlus Kits (KAPA Biosystem, Wilmington, WA) according to the modified manufacturer's instructions.

Detailed protocol is as follows:

5 1. Enzymatic Fragmentation

1. Dilute 500ng input DNA with 10 mM Tris-HCl (pH8.0 - 8.5) in a total of 35 microliters.
2. Assemble each fragmentation reaction on ice by adding the rest of the components in the order shown below:

Component	Volume
dsDNA	35 microliters
KAPA Frag Buffer (10x)	5 microliters
KAPA Frag Enzyme	10 microliters
Total Volume	50 microliters

- 10 3. Vortex gently and spin down briefly. Return the tubes to ice. Proceed immediately to the next step.
4. Incubate in a thermocycler, pre-cooled to 4 degrees Celsius and programmed as below. Set the lid temperature to ≤ 50 degrees Celsius.

Step	Temp	Time
Pre-cool block	4 degrees Celsius	∞
Fragmentation	37 degrees Celsius	8 min
HOLD	4 degrees Celsius	∞

5. Transfer reactions to ice and proceed immediately to the next step.

2. End Repair and A-tailing

- 5 1. In the same tubes, assemble each end repair and A-tailing reaction as follows:

Component	Volume
Fragmented, dsDNA	50 microliters
End Repair & A-tailing Buffer	7 microliters
End Repair & A-tailing Enzyme Mix	3 microliters
Total Volume	60 microliters

2. Vortex gently and spin down briefly.
3. Incubate the tubes at 65 degrees Celsius for 30 min and hold at 4 degrees Celsius. Set the lid temperature to 85 degrees Celsius.

3. Adapter Ligation

1. In the same tubes, assemble each adapter ligation reaction as follows:

Component	Volume
End Repair & A-tailing reaction product	60 microliters
Adapter stock 15uM	5 microliters
PCR grade water	5 microliters
Ligation Buffer	30 microliters
DNA ligase	10 microliters
Total Volume	110 microliters

2. Mix thoroughly and centrifuge briefly, and incubate at 20 degrees Celsius for 30 min.

4. Post-ligation Cleanup

1. In the same tubes, perform a 0.8X SPRI cleanup by combining the following:

Component	Volume
Adapter ligation reaction product	110 microliters
Agencourt AMPure XP reagent	88 microliters
Total Volume	198 microliters

2. Mix thoroughly by pipetting up and down 10 times.
- 5 3. Incubate at room temperature for 10 min to bind DNA to the beads.
4. Place the tubes on a magnet to capture the beads. Incubate for about 5 min till the liquid is clear.
5. Carefully remove and discard the supernatant. Keep the tubes on the magnet and add 200ul of 80% ethanol. Incubate the tubes at RT for
10 ≥ 30 sec and remove the ethanol.
6. Repeat Step 5. Try to remove all residual ethanol without disturbing the beads.
7. Dry the beads at RT for 5 min. Remove the tubes from magnet and resuspend the beads in 55 microliters 10mM Tris-HCl elution buffer.

8. Incubate the tubes at RT for 2 min to elute DNA off the beads.

9. Place the tubes on a magnet to capture the beads and incubate until the liquid is clear. Transfer 50 microliters cleared supernatant to new tubes.

5 **5. Double-size selection**

1. Perform the first 0.6 X size cut to the tubes by adding 30 microliters AMPure XP reagent to eluted 50 microliters DNA library.

2. Mix thoroughly by pipetting up and down 10 times.

3. Incubate at room temperature for 10 min to bind library molecules
10 larger than ~450 bp to the beads.

4. Place the tubes on a magnet to capture the beads. Incubate for about 5 min till the liquid is clear.

5. Transfer 75 microliters of supernatant containing library molecules smaller than ~450bp to new tubes. No beads should be transferred with
15 the supernatant.

6. Perform the second 0.7 X size cut to the tubes by adding 5 microliters AMPure XP reagent.

7. Mix thoroughly by pipetting up and down 10 times.

8. Incubate at room temperature for 10 min to bind library molecules
20 larger than ~300 bp to the beads.

9. Place the tubes on a magnet to capture the beads and incubate till the liquid is clear.

10. Remove and discard the supernatant, which contains library molecules smaller than ~300bp. Keep the tubes on the magnet and add 200 microliters of 80% ethanol. Incubate the tubes at RT for ≥ 30 sec and remove the ethanol.
- 5 11. Repeat Step 5. Try to remove all residual ethanol without disturbing the beads.
12. Dry the beads at RT for 5 min. Remove the tubes from magnet and resuspend the beads in 25 microliters 10mM Tris-HCl elution buffer.
- 10 13. Incubate the tubes at RT for 2 min to elute DNA off the beads.
14. Place the tubes on a magnet to capture the beads and incubate until the liquid is clear. Transfer 20 microliters cleared supernatant to new tubes for PCR amplification.

15

20

6. Library Amplification

1. Assemble each library amplification reaction as follows:

Component	Volume
2X KAPA HiFi HotStart ReadyMix	25 microliters
10X KAPA library amplification primer mix	5 microliters
Adapter ligated library	20 microliters
Total Volume	50 microliters

2. Mix thoroughly and centrifuge briefly.

3. Amplify using the cycling protocol as follows:

Step	Temp	Duration	Cycles
Initial denaturation	98 degrees Celsius	45 sec	1
Denaturation	98 degrees Celsius	15 sec	
Annealing	60 degrees Celsius	30 sec	3-4
Extension	72 degrees Celsius	30 sec	
Final extension	72 degrees Celsius	1 min	1
HOLD	4 degrees Celsius	∞	1

7. Post-Amplification Cleanup

1. In each library amplification tube, perform a 0.7X SPRI cleanup by combining the following:

Component	Volume
Library Amplification reaction product	50 microliters
Agencourt AMPure XP reagent	35 microliters
Total Volume	85 microliters

2. Mix thoroughly by pipetting up and down 10 times.
3. Incubate at room temperature for 10 min to bind DNA to the beads.
4. Place the tubes on a magnet to capture the beads. Incubate for about 5 min till the liquid is clear.
5. Carefully remove and discard the supernatant. Keep the tubes on the magnet and add 200ul of 80% ethanol. Incubate the tubes at RT for ≥ 30 sec and remove the ethanol.
6. Repeat Step 5. Try to remove all residual ethanol without disturbing the beads.

7. Dry the beads at RT for 5 min. Remove the tubes from magnet and resuspend the beads in 35 microliters 10mM Tris-HCl elution buffer.
8. Incubate the tubes at RT for 2 min to elute DNA off the beads.
9. Place the tubes on a magnet to capture the beads and incubate until
5 the liquid is clear. Transfer 30 microliters cleared supernatant to new tubes.
10. HS Qubit will be used for the quantification of DNA library. Analysis of DNA fragments will be achieved by the Agilent 2100 Bioanalyzer.
- 10 11. 4 libraries with equal molarity are pooled and sequenced on NGS platform.

NGS library preparation - RNA

- cDNA library was conducted with KAPA HyperPlus Kits (KAPA Biosystem,
15 Wilmington, WA) according to the modified manufacturer's instructions.
Detailed protocol is as follows:

1. Enzymatic Fragmentation

1. Dilute 1-1000 ng input cDNA with 10 mM Tris-HCl (pH8.0 - 8.5) in a total of 35 microliters.

2. Assemble each fragmentation reaction on ice by adding the rest of the components in the order shown below:

Component	Volume
dsDNA	35 microliters
KAPA Frag Buffer (10x)	5 microliters
KAPA Frag Enzyme	10 microliters
Total Volume	50 microliters

3. Vortex gently and spin down briefly. Return the tubes to ice. Proceed immediately to the next step.

- 5 4. Incubate in a thermocycler, pre-cooled to 4 °C and programmed as below. Set the lid temperature to ≤ 50 °C.

Step	Temp	Time
Pre-cool block	4 degrees Celsius	∞
Fragmentation	25 degrees Celsius	5 min
HOLD	4 degrees Celsius	∞

5. Transfer reactions to ice and proceed immediately to the next step.

3. End Repair and A-tailing

4. In the same tubes, assemble each end repair and A-tailing reaction as follows:

Component	Volume
Fragmented, dsDNA	50 microliters
End Repair & A-tailing Buffer	7 microliters
End Repair & A-tailing Enzyme Mix	3 microliters
Total Volume	60 microliters

5. Vortex gently and spin down briefly.
6. Incubate the tubes at 65 degrees Celsius for 30 min and hold at
5 4 degrees Celsius. Set the lid temperature to 85 degrees Celsius.

3. Adapter Ligation

1. In the same tubes, assemble each adapter ligation reaction as follows:

Component	Volume
End Repair & A-tailing reaction product	60 microliters
Adapter stock 15uM	5 microliters
PCR grade water	5 microliters

Ligation Buffer	30 microliters
DNA ligase	10 microliters
Total Volume	110 microliters

2. Mix thoroughly and centrifuge briefly, and incubate at 20 degrees Celsius for 30 min.

4. Post-ligation Cleanup

- 5 1. In the same tubes, perform a 0.8X SPRI cleanup by combining the following:

Component	Volume
Adapter ligation reaction product	110 microliters
Agencourt AMPure XP reagent	88 microliters
Total Volume	198 microliters

2. Mix thoroughly by pipetting up and down 10 times.
3. Incubate at room temperature for 10 min to bind DNA to the beads.
4. Place the tubes on a magnet to capture the beads. Incubate for about
- 10 5 min till the liquid is clear.

5. Carefully remove and discard the supernatant. Keep the tubes on the magnet and add 200 microliters of 80% ethanol. Incubate the tubes at RT for ≥ 30 sec and remove the ethanol.
6. Repeat Step 5. Try to remove all residual ethanol without disturbing the beads.
7. Dry the beads at RT for 5 min. Remove the tubes from magnet and resuspend the beads in 23 microliters 10mM Tris-HCl elution buffer.
8. Incubate the tubes at RT for 2 min to elute DNA off the beads.
9. Place the tubes on a magnet to capture the beads and incubate until the liquid is clear. Transfer 20 microliters cleared supernatant to new tubes.

5. Library Amplification

1. Assemble each library amplification reaction as follows:

Component	Volume
2X KAPA HiFi HotStart ReadyMix	25 microliters
10X KAPA library amplification primer mix	5 microliters
Adapter ligated library	20 microliters

Total Volume	50 microliters
--------------	-----------------------

2. Mix thoroughly and centrifuge briefly.

3. Amplify using the cycling protocol as follows:

Step	Temp	Duration	Cycles
Initial denaturation	98 degrees Celsius	45 sec	1
Denaturation	98 degrees Celsius	15 sec	
Annealing	60 degrees Celsius	30 sec	5-9
Extension	72 degrees Celsius	30 sec	
Final extension	72 degrees Celsius	1 min	1
HOLD	4 degrees Celsius	∞	1

6. Post-Amplification Cleanup

- 5 1. In each library amplification tube, perform SPRI cleanup twice by combining the following:

Component	First cleanup	Second cleanup
Library Amplification reaction product	50 microliters	50 microliters
Agencourt AMPure XP reagent	40 microliters	35 microliters
Total Volume	90 microliters	85 microliters

2. Mix thoroughly by pipetting up and down 10 times.
3. Incubate at room temperature for 10 min to bind DNA to the beads.
4. Place the tubes on a magnet to capture the beads. Incubate for about 5 min till the liquid is clear.
5. Carefully remove and discard the supernatant. Keep the tubes on the magnet and add 200ul of 80% ethanol. Incubate the tubes at RT for ≥ 30 sec and remove the ethanol.
6. Repeat Step 5. Try to remove all residual ethanol without disturbing the beads.
7. Dry the beads at RT for 5 min. Remove the tubes from magnet and resuspend the beads in 53 microliters 10mM Tris-HCl elution buffer.
8. Incubate the tubes at RT for 2 min to elute DNA off the beads.
9. Place the tubes on a magnet to capture the beads and incubate until the liquid is clear. Transfer 50 microliters cleared supernatant to new tubes.
10. Repeat steps 1-9 with second cleanup. In second cleanup, resuspend beads with 23 microliters 10mM Tris-HCl elution buffer (step 7) and Transfer 20 microliters cleared supernatant to new tubes (step 9).

11. HS Qubit will be used for the quantification of DNA library.

Analysis of DNA fragments will be achieved by the Agilent 2100 Bioanalyzer.

12. 4 libraries with equal molarity are pooled and sequenced on NGS platform.

5

CLAIMS

What is claimed is:

1. An experimental pipeline for pan-domain species nucleotide extraction and next generation sequencing library preparation, comprising:
 - 5 a) preparing DNA/RNA genome libraries from collected metagenome samples at a single bioparticle sensitivity,
wherein a radiation-based reagent-decontamination protocol is implemented to remove any traces of amplifiable nucleotides from reagents,
 - 10 wherein a sterilized hydrophobic or hydrophilic filter membrane is used to collect nano-gram-level amount of airborne biological particulates at the minimum in between 12 to 72 hours,
wherein, with an amount range above 1 ng of nucleotides of airborne biological samples, extracting DNA and RNA concurrently
15 from the collected samples of any nature by a chemistry/physics-based pipeline at a single bacterial particle sensitivity,
wherein a non-PCR-based linear amplification technique is applied for DNA and RNA after converted to complementary DNA,
20 and

- wherein, with limited RNAs, depleting most ribosome RNA with selective primers for preferential enrichment of non-ribosome RNA sequences; and
- (b) using cocktail enzyme to fragmentize DNA/cDNA generated to minimize the loss of nucleotide materials.
- 5
2. A pan-domain species detection bioinformatics pipeline for next generation sequencing data, comprising:
- (a) building a non-redundant genome sequence database incorporating publicly available genome sequences and/or select organisms of interest;
- 10
- (b) building a taxonomy database to store taxonomy information for every sequence in the database;
- (c) creating de-duplicated sequencing reads in fastq format by removing exact paired-duplicated reads from raw sequencing reads;
- 15
- (d) trimming the de-duplicated sequencing reads;
- (e) mapping the trimmed de-duplicated sequencing reads to a human reference genome therewith creating trimmed de-duplicated non-human reads in fastq format;
- 20
- (f) creating assembled contigs from the trimmed de-duplicated non-human reads in fastq format by a rigorous reads assembly method with specific parameters to minimize chimeric assembly;

(g) classifying the assembled contigs by searching against the non-redundant genome sequence database; and

(h) outputting the classified results,

wherein the pipeline steps (a)-(h) are executed by a computer or one or

5 more computer processors.

1/13**Collection**

1. Device is turned on and actively collecting samples via air pump

Sample processing and analysis

2a. Filter is removed from device for biotic materials processing

3a. DNA and RNA extraction from the filter

4a. DNA and RNA amplification and library prep

5a. Sequencing on Next Generation Sequencing (NGS) platform

6a. Sequence analysis and identification of biotic agents via LCA

2b. Resin-containing cartridge is removed from device for abiotic materials processing

3b. Compound extraction using methanol

4b. LC/MS analysis using UPLC-coupled exactive orbitrap mass spectrometer

5b. Raw LC/MS files are analyzed using PAVA, XCMS and SIMCA softwares

Results and interpretation

Reporting of biotic and abiotic materials that are present in the air

FIG. 1

2/13

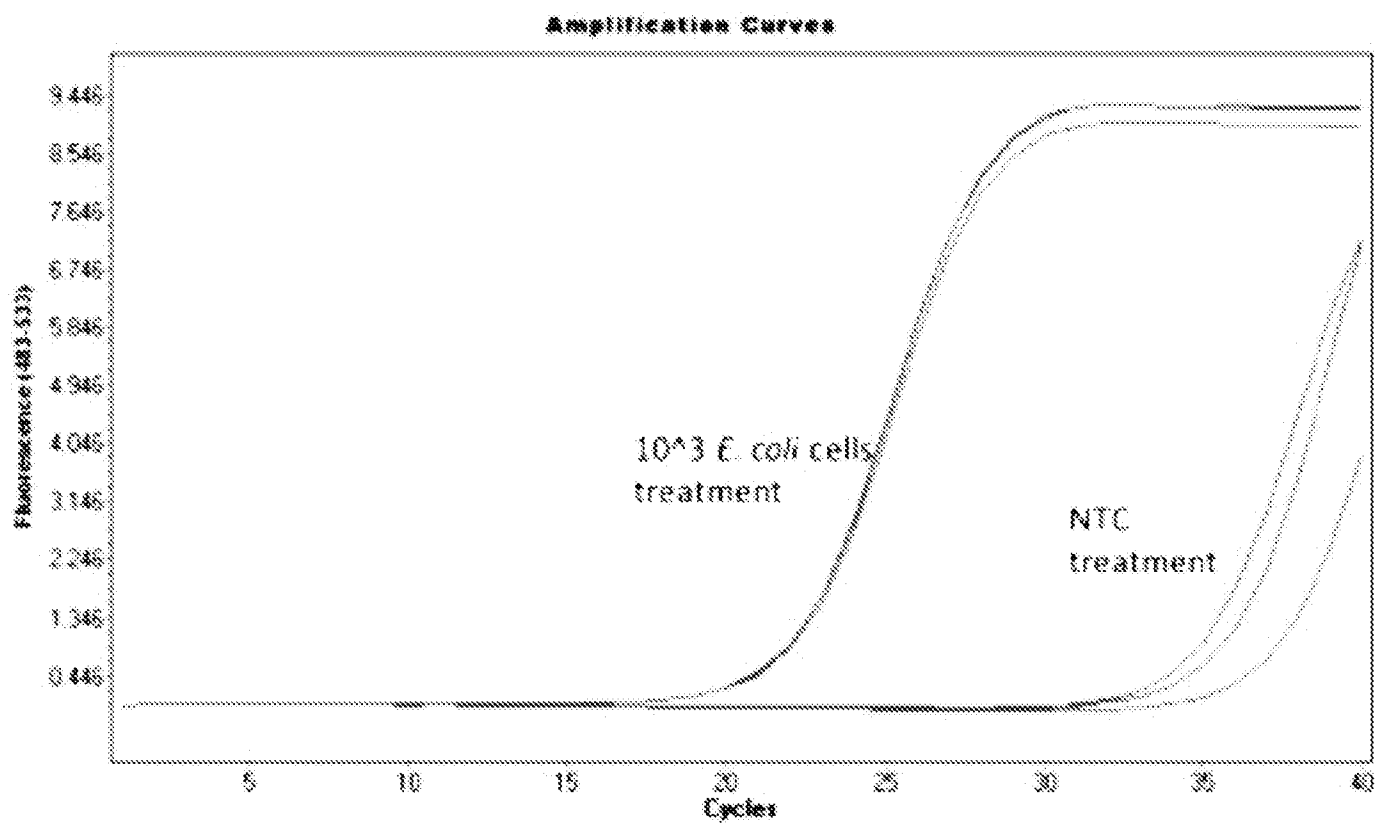
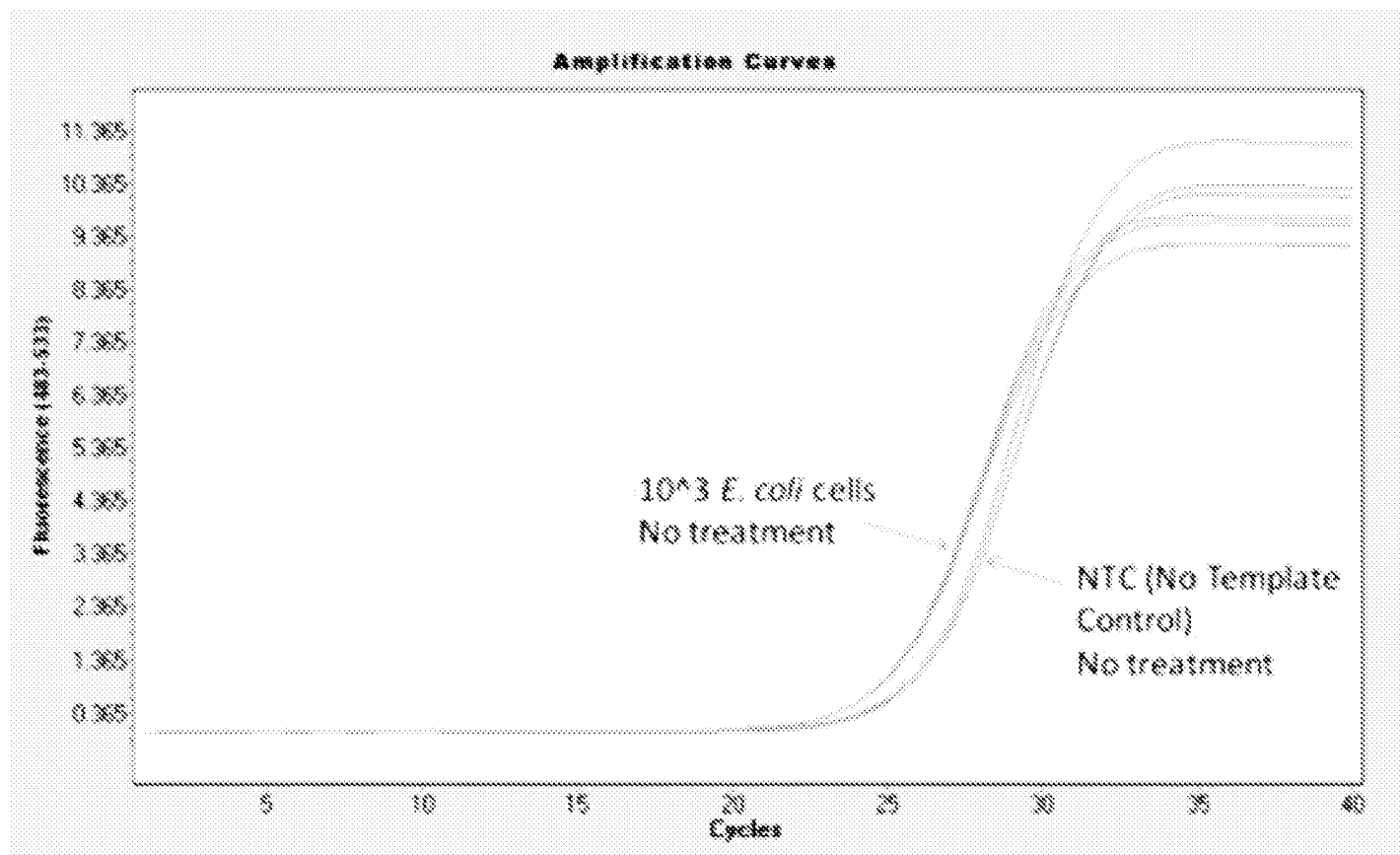


FIG. 2

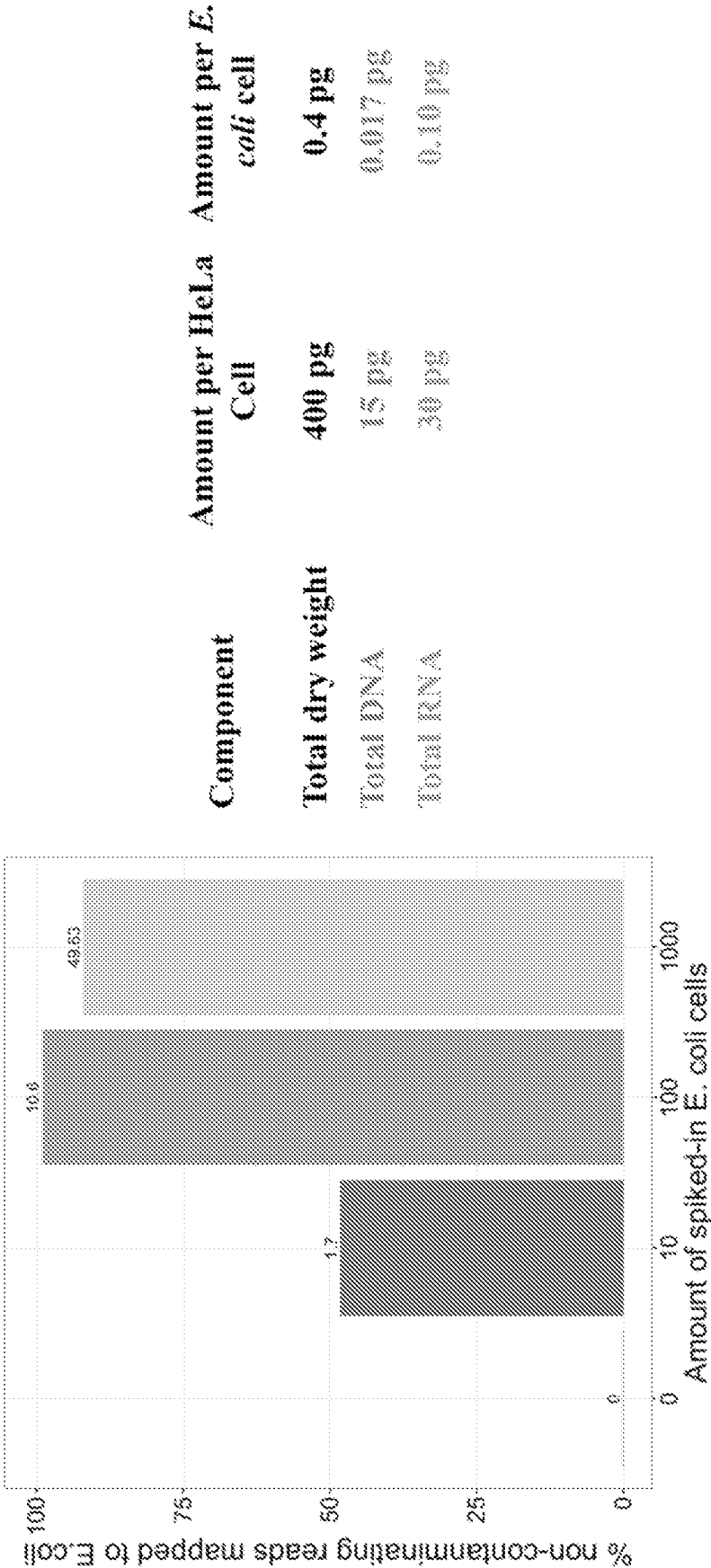
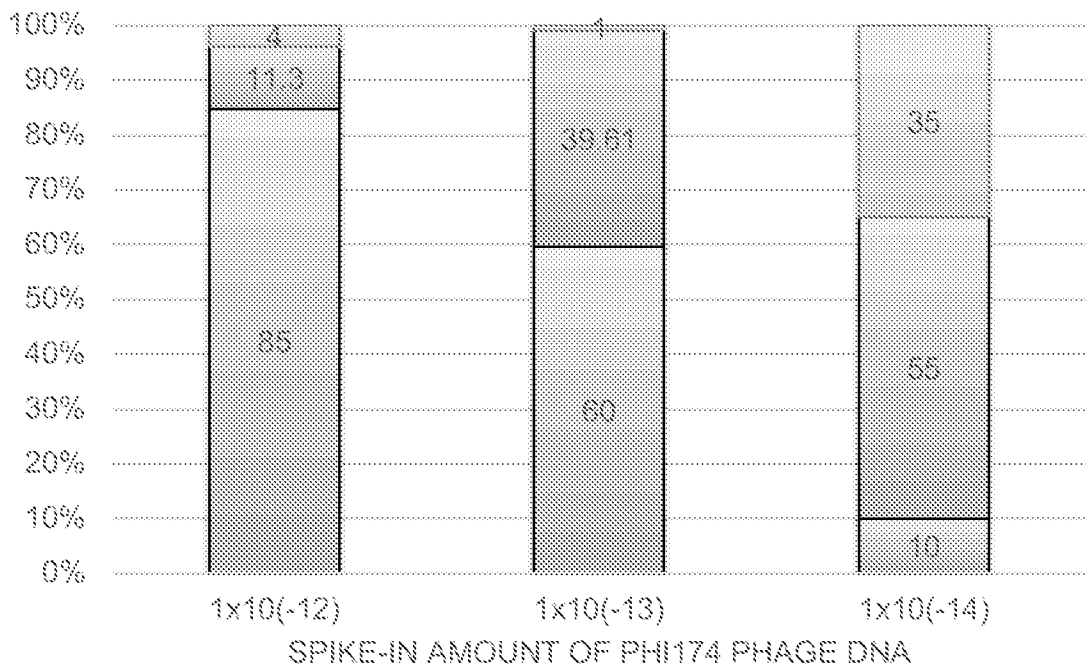


FIG. 3

4/13

Spike-in results of sample 1



From bottom
to top

5 pg 1000 E.coli cell phi174 human others 50 fg 10 E.coli cell

Spike-in results of sample 2

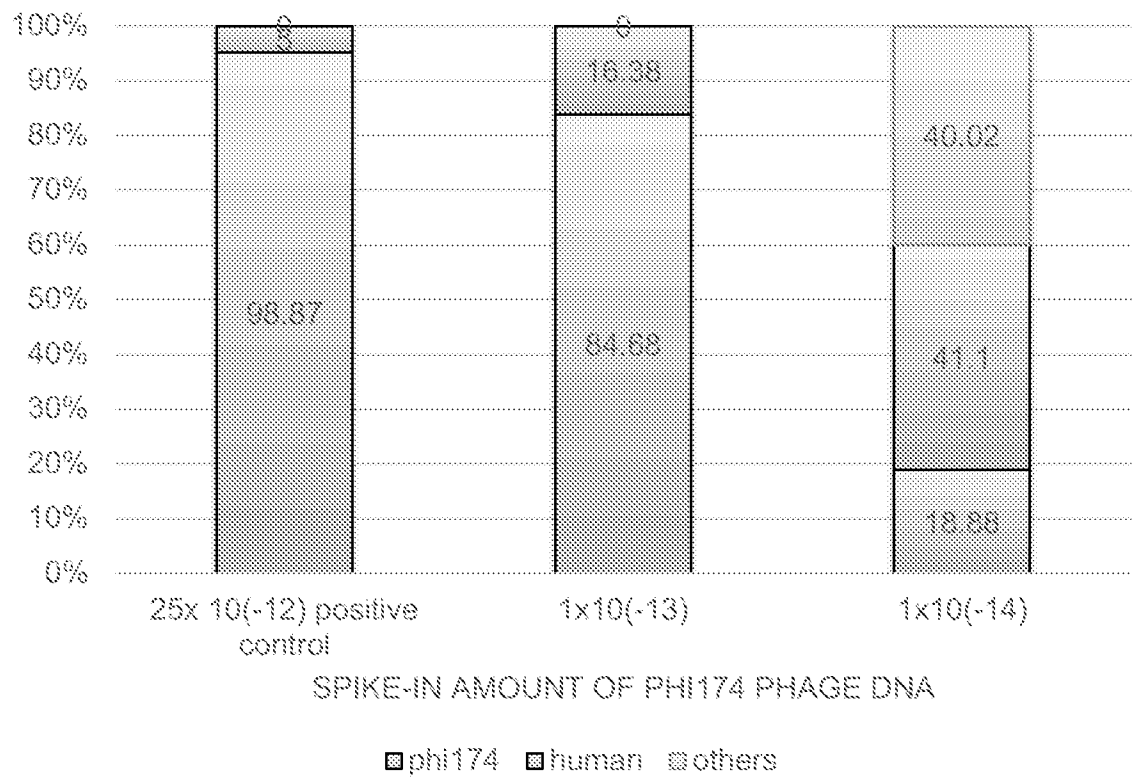
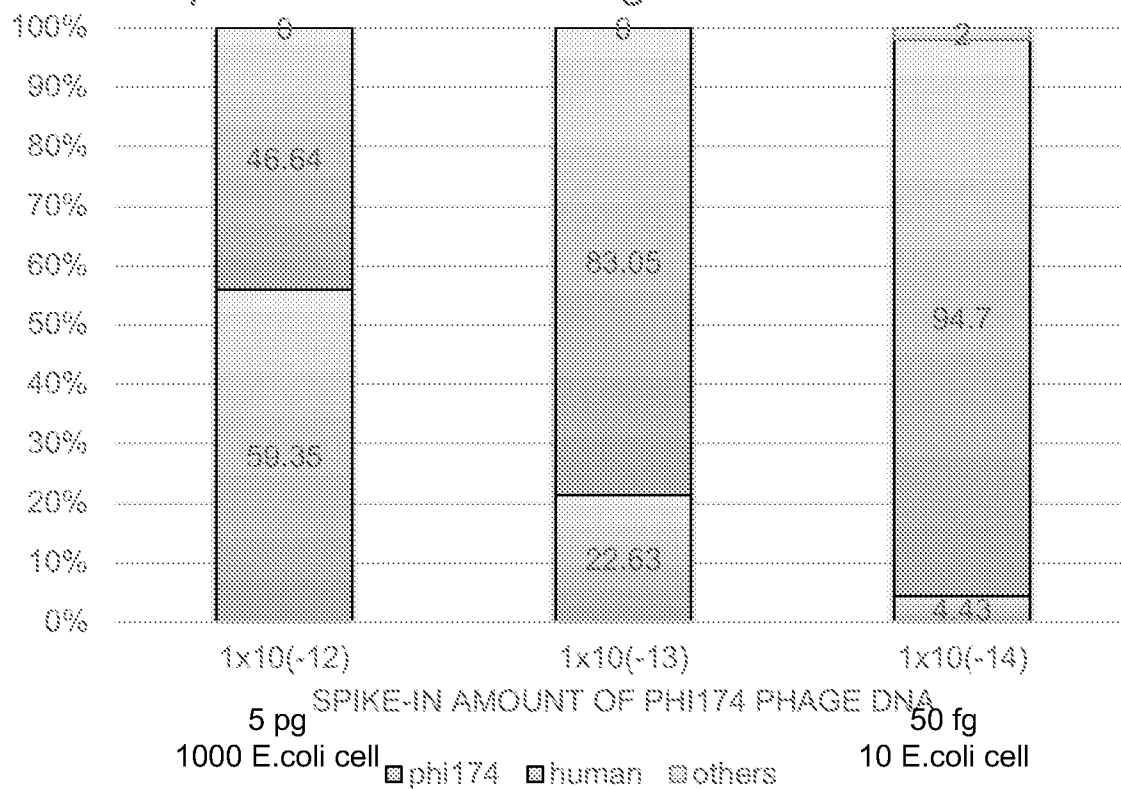
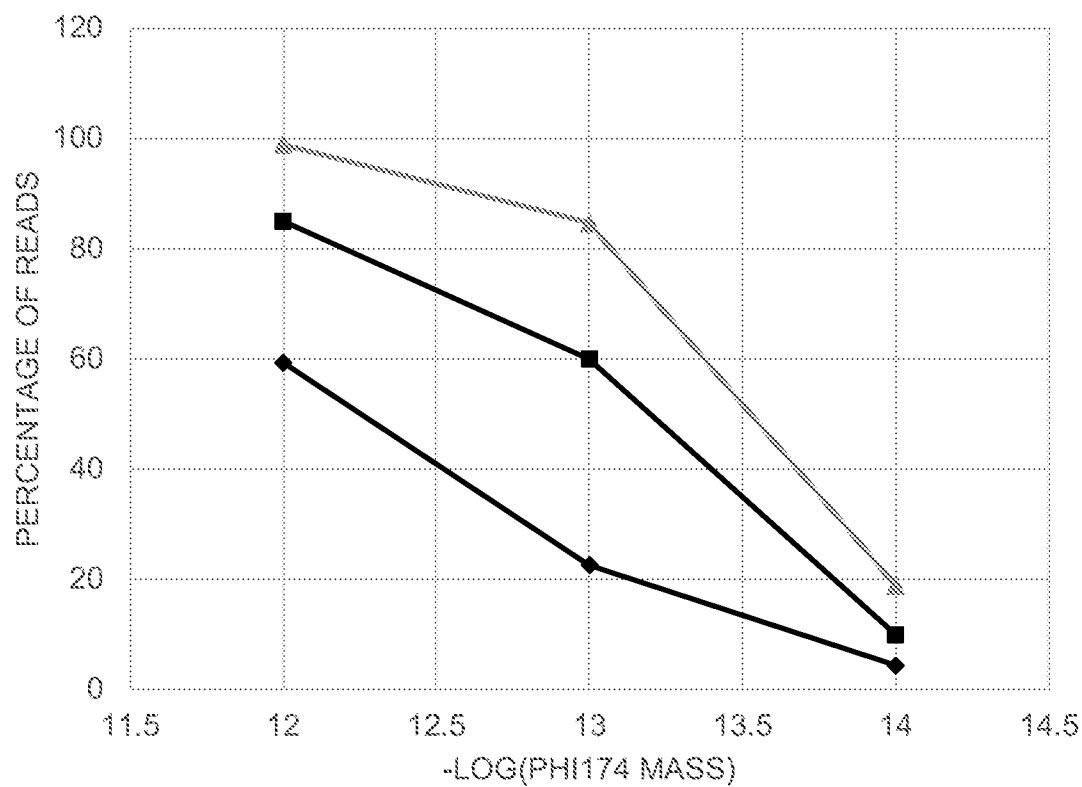


FIG. 4A

5/13**Spike-in results of Negative Control****PHI174 SPIKE-IN AMOUNT VS READS
MAPPED TO PHI174****FIG. 4B**

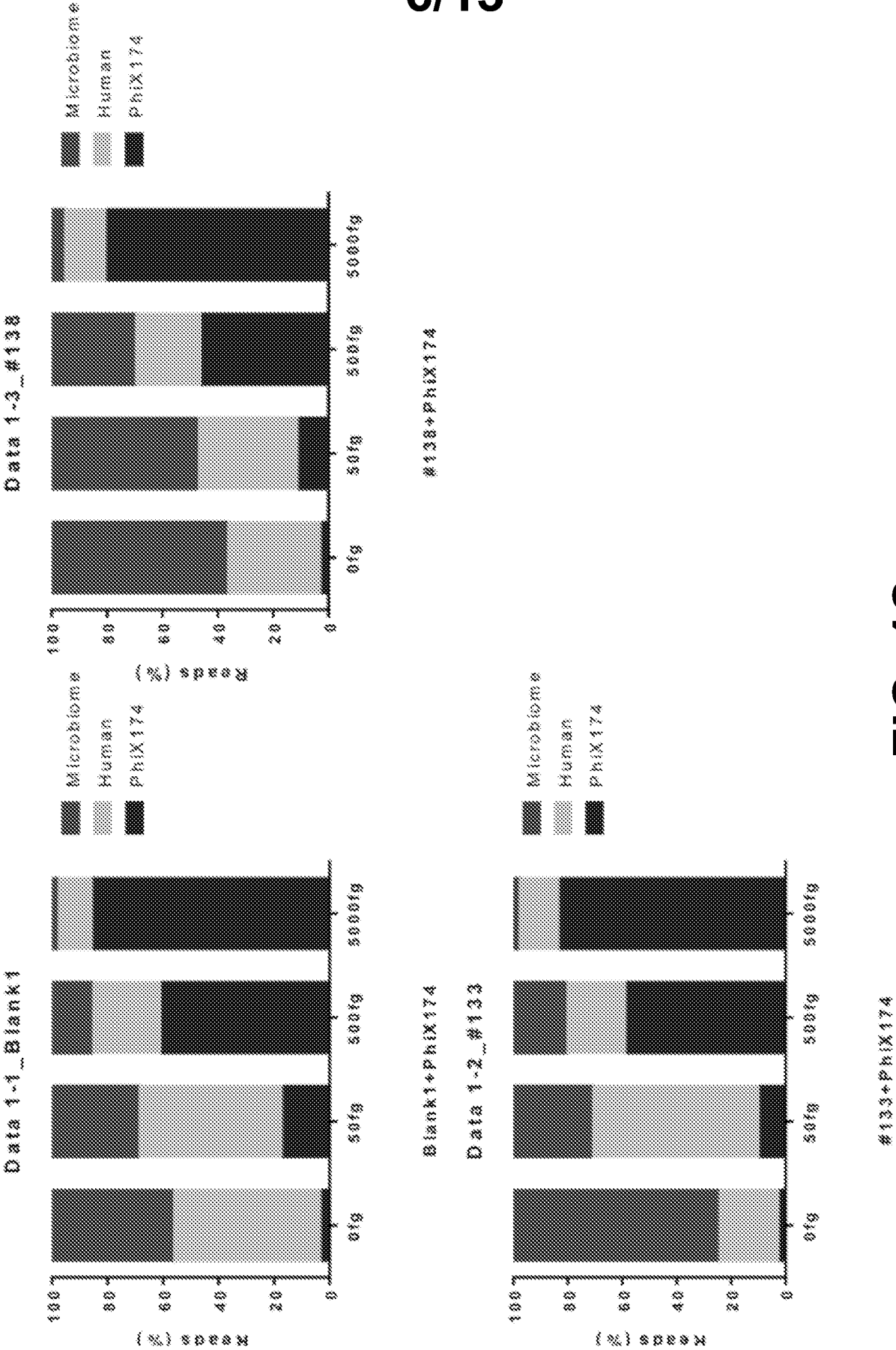


FIG. 4C

7/13

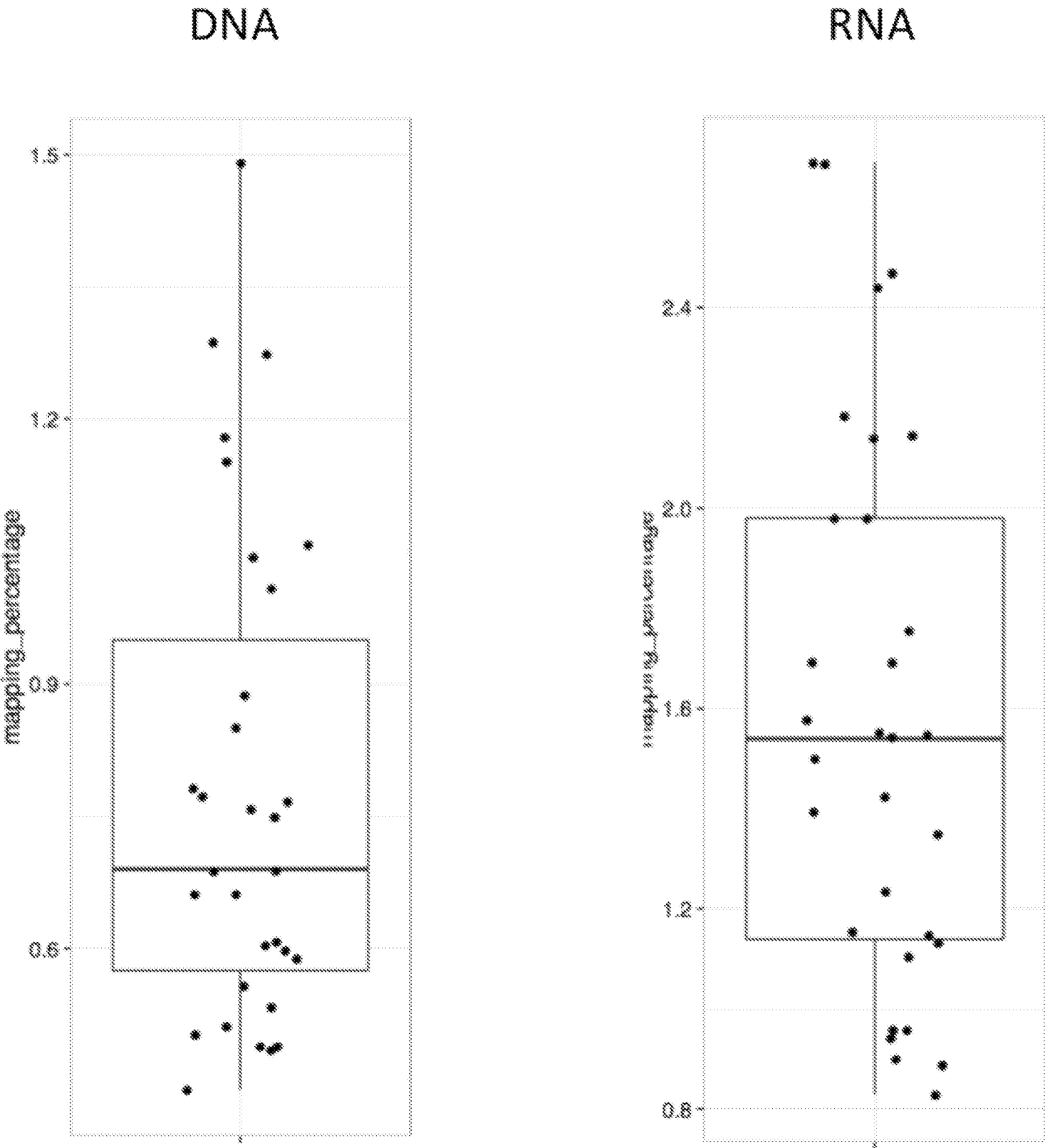
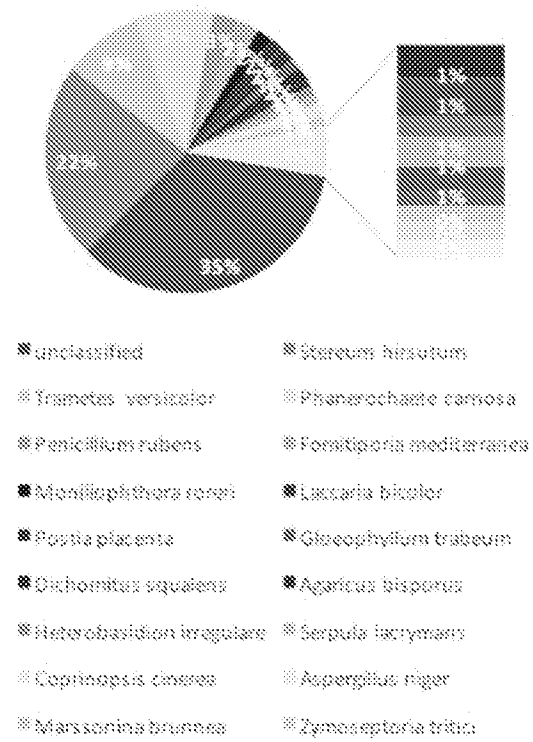


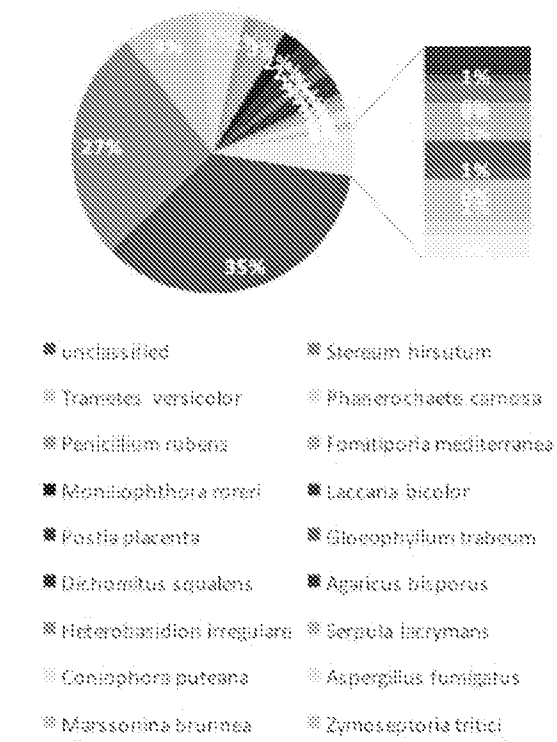
FIG. 5

8/13

Filter #1 by species



Filter #2 by species



CORRELATION BETWEEN REPLICATE FILTERS PEARSON R = 0.9085

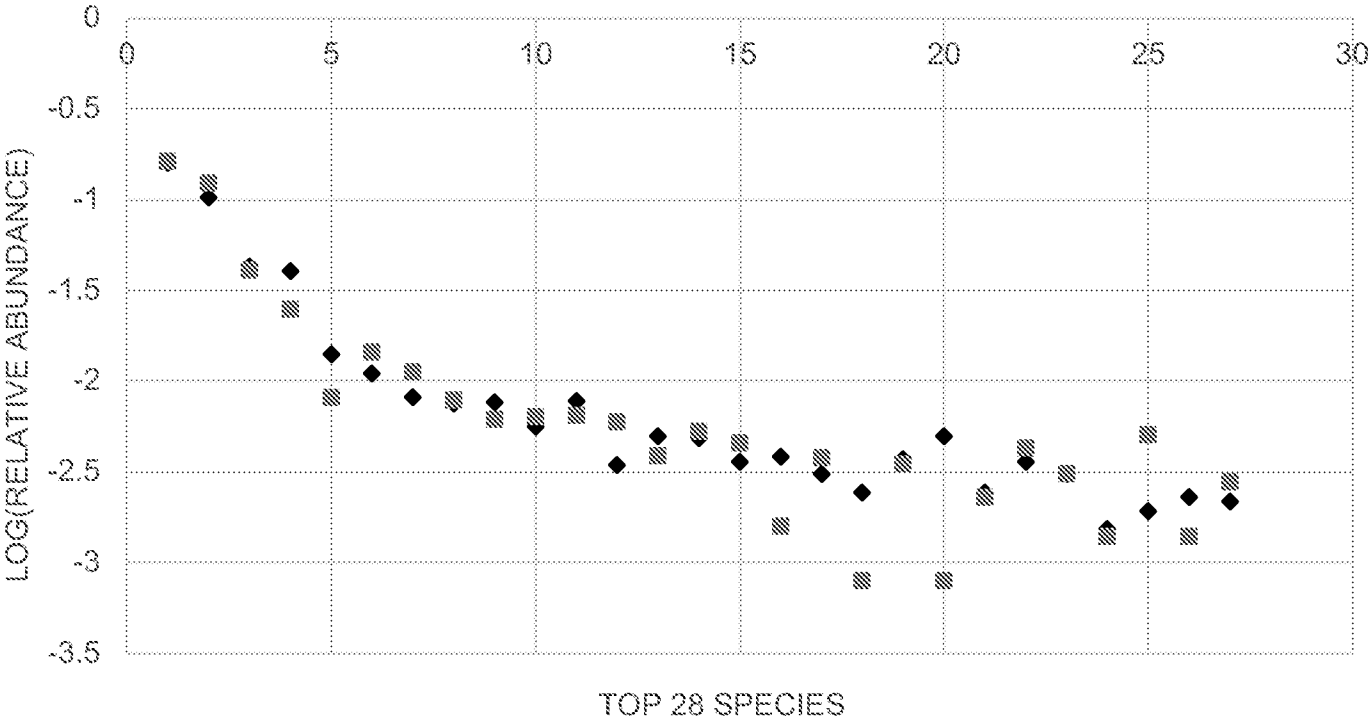


FIG. 6

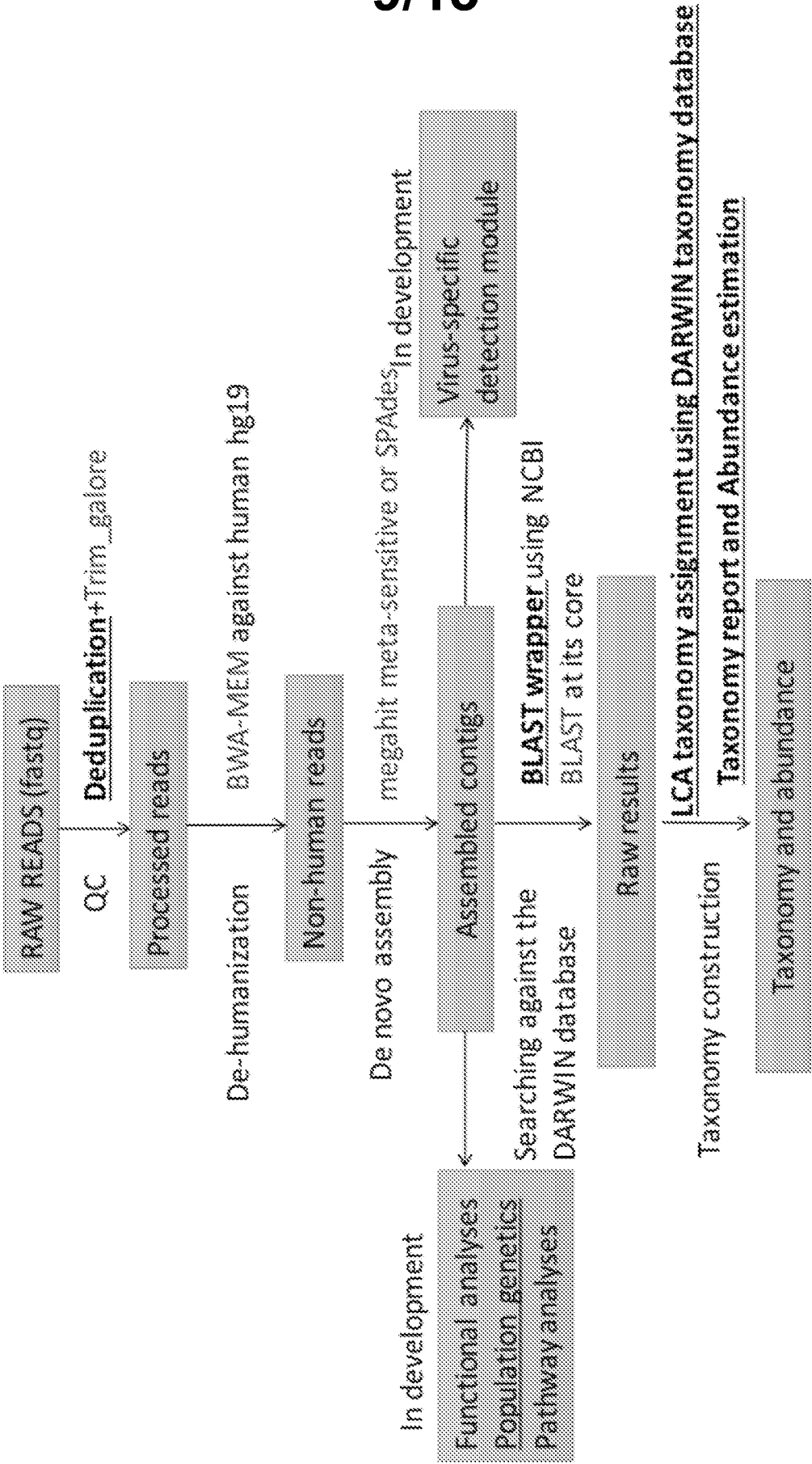
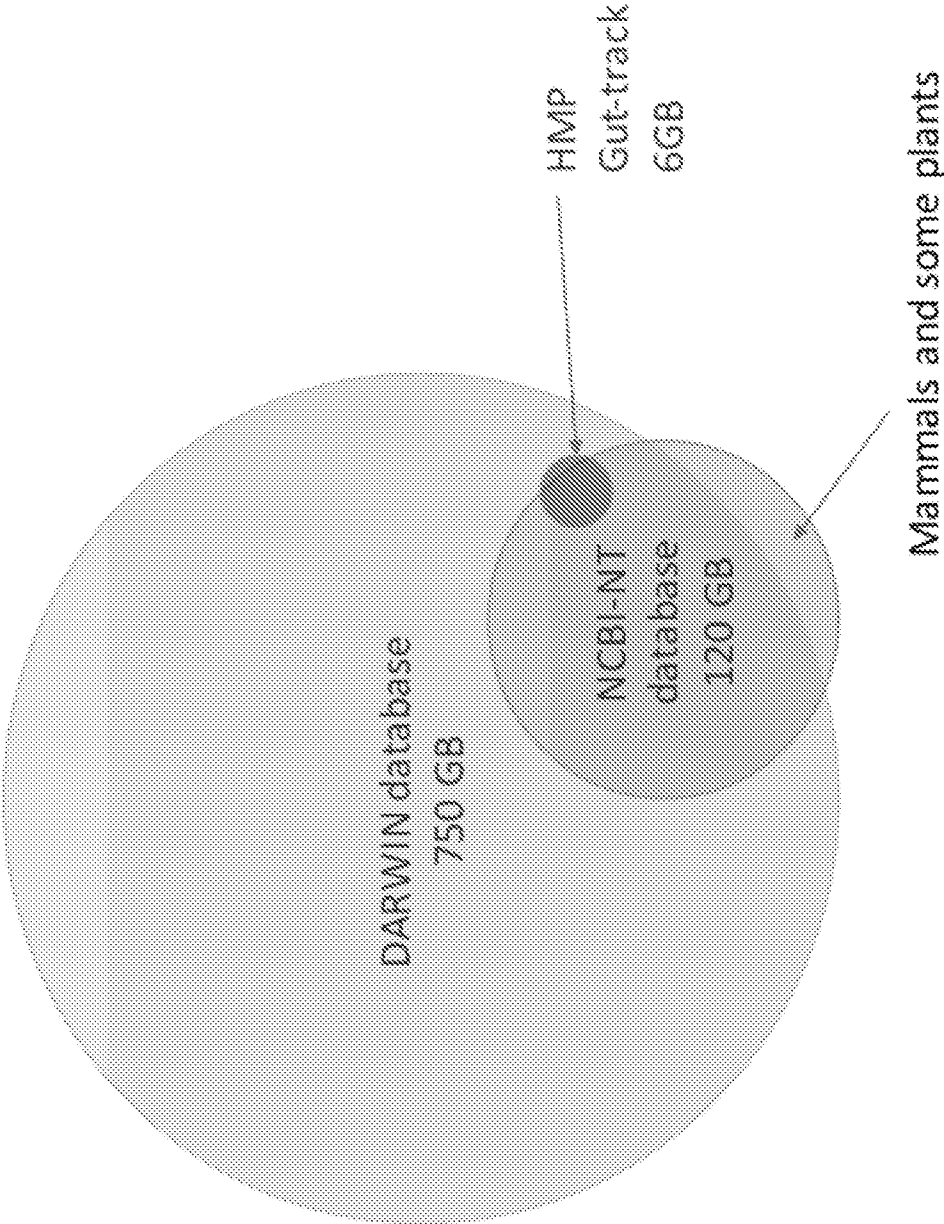


FIG. 7

10/13



Bacteria, Fungi, Plants, Viruses,
Archaea, simple eukaryotes,
Invertebrates, Vertebrates.

82 million entries,
40,000 species
99.97% index rate

FIG. 8

11/13

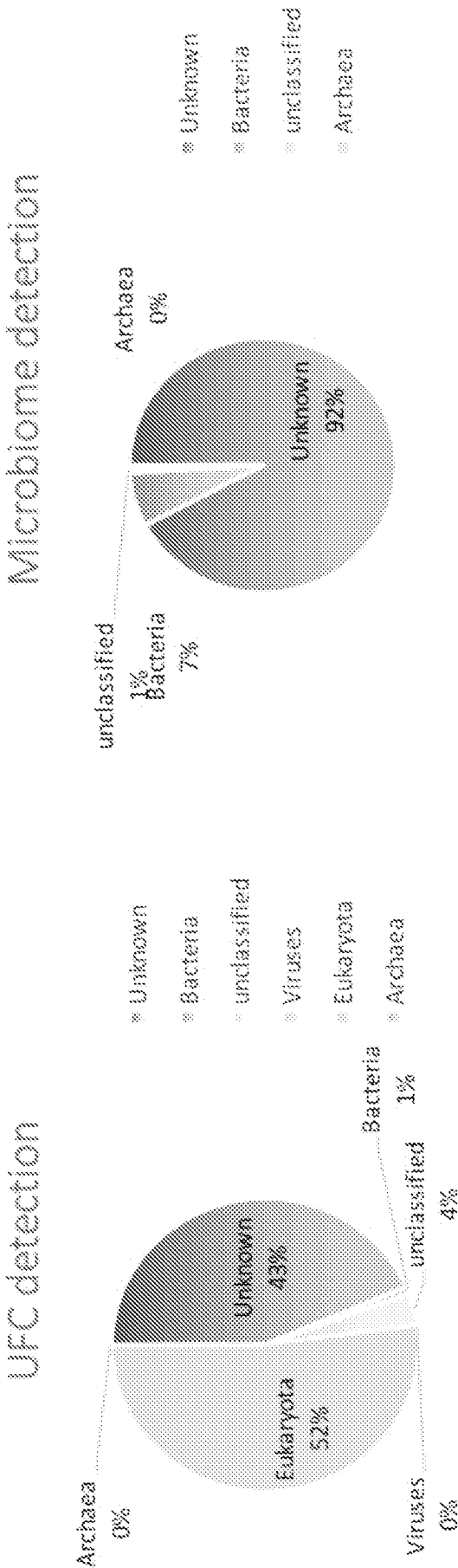


FIG. 9

12/13

Copies	Name	detection
200	Human parainfluenza virus 3	YES
200	Rhinovirus 1A	YES
400	Human metapneumovirus	YES
400	Betacoronavirus 1 (coronavirus OC43)	YES
400	Enterovirus B	YES
1000	Human respiratory syncytial virus	YES
1000	Coronavirus 229E (alphacoronavirus)	YES
1000	Human parainfluenza virus 2	YES
1000	Influenza B virus	YES
1400	Influenza A virus	YES
10000	Human immunodeficiency virus 1	YES
10000	Cowpea mosaic virus	YES

FIG. 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2018/028542

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12N 15/10; G06F 19/14; G06F 19/24 (2018.01)

CPC - C12N 15/1017; C12N 15/1003; G06F 19/14; G06F 19/24 (2018.05)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

USPC - 435/6.11; 506/16; 702/20 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2017/053446 A2 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA) 30 March 2017 (30.03.2017) entire document	2
--		--
Y		1
Y	US 2007/0042422 A1 (HOEFT et al) 22 February 2007 (22.02.2007) entire document	1
Y	GRIFFIN et al. "Observations on the use of membrane filtration and liquid impingement to collect airborne microorganisms in various atmospheric environments," Aerobiologia, 29 June 2010 (29.06.2010), Vol. 27, Iss. 1, Pgs. 25-35. entire document	1
Y	US 2012/0316074 A1 (SAXONOV et al) 13 December 2012 (13.12.2012) entire document	1
P, X	WO 2018/045315 A1 (WISCONSIN ALUMNI RESEARCH FOUNDATION) 08 March 2018 (08.03.2018) entire document	1, 2
A	US 2013/0157874 A1 (DOWD) 20 June 2013 (20.06.2013) entire document	1, 2
A	LEIPZIG. "A review of bioinformatic pipeline frameworks," Briefings in Bioinformatics, 24 March 2016 (24.03.2016), Vol. 18, No. 3, Pgs. 530-536. entire document	1, 2

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

27 June 2018

Date of mailing of the international search report

13 JUL 2018

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774