Journal of Applied Microbiology

The selection of shotgun metagenomics software introduces biases in microbial profiling and pathogen detection

**Abbreviated running headline**

Metagenomics software selection biases

**Author names**

Ruijie Xu1,2, Sreekumari Rajeev3,†,\*, Liliana C. M. Salvador1,2,4,†,\*

**Affiliation**

1Institute of Bioinformatics, University of Georgia, Athens, GA, 30602, USA

2Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA, 30602, USA

3Department of Biomedical and Diagnostic Sciences, College of Veterinary Medicine, University of Tennessee, Knoxville, TN, 37996, USA

4Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, 30602, USA

† These authors contributed equally to this work

**\*Corresponding authors**

Liliana C. M. Salvador: Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, 501 D. W. Brooks Drive, Athens, GA 30602, Email: [salvador@uga.edu](mailto:salvador@uga.edu)

Sreekumari Rajeev: Department of Biomedical and Diagnostic Science, College of Veterinary Medicine, University of Tennessee, 2407 River Drive, Knoxville, TN 37996, Email: srajeev@utk.edu

**Abstract**

**Aim**

Shotgun metagenomic sequencing analysis is widely used for microbial profiling of biological specimens and pathogen detection. However, very little is known about the technical biases caused by the choice of analysis databases and software. In this study, we evaluated popular shotgun metagenomics taxonomical profiling software to characterize the microbial compositions of wildly collected rodent samples.

**Method and Results**

We analyzed shotgun metagenomic sequence data from three sets of wild rodent tissue samples collected from St.Kitts using the four different databases and nine most widely used metagenomics software. We demonstrated the discrepancies in results when different databases and software were used, which cause significant differences in microbial communintiy characterizations. Our analysis also showed that these software differed in their ability to identify the presence of *Leptospira,* a major zoonotic pathogen of one health importance in comparison to traditional methods.

**Conclusions**

Significant differences in compositional profiles for the same dataset while using different databases and software may lead to divergent biological conclusions in microbial profiling and zoonotic pathogens detection.

**Significance and Impact of Study**

This study highlights the importance to warrant caution while using shotgun metagenomics for pathogen detection and interpretation of taxonomical profiling analyses.

**Keywords**

shotgun metagenomic sequencing, pathogen detection, *Leptospira*, next-generation sequencing

**Introduction**

Studies analyzing the composition of microbial communities have been utilized in diverse study fields, such as ecology (Galbraith *et al.*, 2018; Grossart *et al.*, 2020), agriculture (Mashiane *et al.*, 2017; Granjou and Phillips, 2019), human and animal health (Chen *et al.*, 2019; Zhong *et al.*, 2019), and pharmacology (Chavira *et al.*, 2019; Wang *et al.*, 2019), as well as in zoonotic agent detection (Tun *et al.*, 2012). Zoonotic origin pathogens are responsible for over 60% of the infectious diseases identified in humans and can cause significant social and economic burdens (Karesh *et al.*, 2012; Cuervo-Soto, López-Pazos and Batista-García, 2018). Traditional methods used to identify the microbial agents within a biological specimen have relied on different laboratory techniques, including culture (Handelsman, 2004), antigen detection (Desmonts and Remington, 1980; Lequin, 2005), and nucleic acid marker detection (Yang and Rothman, 2004; Driscoll, 2009) protocols. However, these laboratory methods are limited to studying a single pathogen of interest and lack the ability to scrutinize the community of microorganisms potentially present in a sample. Next-Generation Sequencing (NGS) technologies have provided researchers with a set of culture-independent tools that identify pathogens directly from DNA sequences (Ghosh, Mehta and Khan, 2019), and characterize the diversity and abundance of microbial populations in biological specimens. These characteristics have led to the emergence of NGS technologies as popular tools for microbial profiling and pathogen detection (Tun *et al.*, 2012; Skarżyńska *et al.*, 2020; Grützke *et al.*, 2021).

Taxonomical profiling analysis in the metagenomics discipline utilizes two popular approaches: the 16S rRNA and the shotgun metagenomic sequencing-based approach (Jovel *et al.*, 2016). The 16S rRNA sequencing-based method uses polymerase chain reaction (PCR) to amplify hypervariable regions of bacterial 16S rRNA gene and compares these regions to a 16S reference database (DB) (Johnson *et al.*, 2019). In contrast, the shotgun metagenomic sequencing-based approach sequences all given DNA present in a sample (Sharpton, 2014). Although lower in cost (Breitwieser, Lu and Salzberg, 2019), 16S rRNA markers are only available in the genomes of most bacteria and archaea and suffers from primer amplification biases (Woese, Kandlert and Wheelis, 1990; Janda and Abbott, 2007). On the other hand, the taxonomical profiling of shotgun metagenomics sequencing data is done by comparison with a reference whole-genome DB. Since the data contain all genetic information present in the sample, this approach avoids the amplification biases observed in 16S rRNA sequencing (Fouhy *et al.*, 2016; Ranjan *et al.*, 2016) and increase the resolution of microbial identification (Durazzi *et al.*, 2021). Most importantly, it has broader applications such as functional profiling and allows for the identification of viruses and other microorganisms with simple genomes (Clark and Pazdernik, 2016).

Currently developed shotgun metagenome sequencing-based taxonomical profiling software can be separated into two groups, the alignment-based software and the alignment free software. The alignment-based software, including BLASTN (Altschul *et al.*, 1990; Johnson *et al.*, 2008; Camacho *et al.*, 2009), which aligns sequences at nucleotide level, and Diamond (Buchfink, Xie and Huson, 2015), which aligns at the protein level, has been used as the standard for metagenomics profiling due to their high sensitivity . With the high sensitivity, these alignment-based software also suffers from a large trade-off in the time and computational resources they require to build alignement for the amount of sequences involved in metagenomics profiling (Cannings, 2004; Zielezinski *et al.*, 2017). Furthermore, recent investigations in alignment-based methods has also reported alignment-based software’s decrease in sensitivity with more mosaic genomes (ex. viruses) (Zielezinski *et al.*, 2017). With response to the downsides of alignment-based software, multiple software were developed in the recent years with alignment-free algorithms. Some of these software, represented by Kraken2 (Wood, Lu and Langmead, 2019, p. 2) and CLARK (Ounit *et al.*, 2015), were designed with the k-mer matching algorithms, where only subtrings of sequences were matched (Healy and Chambers, 2014). Some software, represented by Metaphlan3 (Truong *et al.*, 2015; Beghini *et al.*, 2021), was designed to identify the unique genetic markers within each microbial taxon. Other software, such as Centrifuge (Kim *et al.*, 2016) and Kaiju (Menzel, Ng and Krogh, 2016), optimizes the time and resources of profiling by compressing the reference microbial genomes into the index structures for storaging and searching (Burrows and Wheeler, 1994), at the nucleotide and protein level, respectively. In addition to the software mentioned above, some software were developed to improve the results of the other software, for example, Bracken (Lu *et al.*, 2017) was developed to improve Kraken2’s result by eliminating the false positive assignments using a Bayesian framework, and CLARK-s (Ounit and Lonardi, 2016) was designed to improve the sensitivity of CLARK with the use of spaced Kmers. Previous benchmarks on shotgun metagenomic sequencing taxonomical profiling software have evaluated the performances these software using either in silico or in vitro datasets (Peabody *et al.*, 2015; Escobar-Zepeda *et al.*, 2018; Ye *et al.*, 2019a). The advantage of using these artificial datasets is that their performances can be evaluated by comparing their microbial profiles with the known composition of the artificial datasets. However, the performance of these software to analyze the microbial profiling and diagnostic applications of biological specimens has been less studied. For samples collected from wild animals, the microbiome compositions are unknown and potentially contain taxa that do not have genomes available in the reference DB . These situations can become a potential source of technical errors for accurate detection and profiling a sample's microbiome.

In this study, we compared the microbial profiles of tissue samples from two species of *Rattus* (*Rattus rattus* and *Rattus norvegicus*) using four different DBs and nine different software metioned above. We specifically address how the use of different DBs and software influence the profiling of the rat samples, and how these influences can affect the downstream analyses. We also focused on the specific detection of the zoonotic pathogen *Leptospira* in rat kidneys. The objectives of the current study are to 1) compare the taxonomical profiles classified by the four Kraken2 DBs ; 2) compare the microbial profiles of our dataset classified by the nine metagenomics profiling software; 3) identify the presence of potential zoonotic pathogens such as *Leptospira* from each software’s profiling results; 4) address if different software can bias the indices characterizing within samples microbial diversity and between samples microbial relationships; and 5) compare taxa identified significantly different in abundance between different tissue samples from each software’s microbial profile. We present data demonstrating the significant differences among the characterizations of the microbial communities analyzed from the microbial profiles obtained using different DB and software. Through comparisons, we have demonstrated the profiling results of Metaphlan3 were least sensitive overall in profiling the microbial communities of the rat samples. BLASTN, CLARK, CLARK-s, and Kaiju were more sensitive in identify the presence of viruses, although the identity of virus taxa identified by these software were not consistent. For microbial community characterization, within-samples’ microbial richness were largely biased by software selection, but less with the DB selection, but the evenness of the samples were reported more consistently across DBs and software. Relationships between communities reported by the classifications of different software were also large biased, but the most distinctive relationships within the rat samples were able to reported by all software except for Metaphlan3. Kaiju and Centrifuge were most sensitive in differentially abundant (DA) taxa identification, especially with virus and archaea taxa, while Diamond were less sensitive in this analysis compare to other software. As for *Leptospira* diagnosis, Centrifuge was the most sensitive software, which reported the presence of the bacteria in sample not detected by any other software or by tradtional techniques. This study presents the biases introduced by metagenomic profiling software for microbial community characterization and the limit of using shotgun metagenomics as the tool for pathogen detection.

**Materials and Methods**

**Samples.** Tissue samples from the kidney (K), spleen (S), and lung (L) were obtained from four rats from two different species, *Rattus rattus* (R28) and *Rattus* *norvegicus* (R22, R26, and R27). Rats were captured from the island of Saint Kitts (longitude 17.3434° N and latitude – 62.7559°W) following protocols approved by the Ross University School of Veterinary Medicine (RUSVM) IACUC (approval # 17-01-04). DNA was extracted from samples using DNeasy Blood and Tissue Kits (QIAGEN Scientific Inc., MD, USA), following the manufacturer's protocol.

**Metagenomic shotgun sequencing.** DNA sample quality was assessed via analysis of the DNA purity and integrity with the agarose gel. DNA purity (OD260/OD280) and concentration were measured using the Nanodrop and Qubit 2.0. The library for metagenomic sequences was constructed with 1 μg DNA per sample. Sequencing libraries were generated using NEBNext® Ultra™ DNA Library Prep Kit for Illumina following manufacturer's instructions. The DNA sample was fragmented (350 bp), end-polished, A-tailed, ligated with Illumina sequencing adaptor and amplified with the PCR technique. The PCR products were then purified for sequencing. Before sequencing, samples were clustered on a cBot Cluster Generation System, then sequenced on an Illumina HiSeq platform for paired-end reads.

**Data pre-processing**. Sequencing adapters, low-quality reads, and host DNA reads within the metagenomic samples were removed using the software KneadData (The Huttenhower Lab, no date) with the default Trimmomatic (Bolger, Lohse and Usadel, 2014) (version 0.33) settings (SLIDINGWINDOW:4:20 MINLEN:50) and the "—very-sensitive" Bowtie (Langmead *et al.*, 2019) (version 2.3) option. The hosts' reference sequences, which were used to separate host reads from the microbial reads, were downloaded from the NCBI's RefSeq DB (Human: GCA\_000001405.28\_GRCh38.p13; *R. norvegicus*: GCF\_015227675.2\_mRatBN7.2; *R. rattus*: GCF\_011064425.1\_Rrattus\_CSIRO\_v1).

**Metagenomic profiling.**

Software. Nine software (BLASTN, Diamond, Kraken2, Bracken, Centrifuge, CLARK, CLARK-s, Metaphlan3, Kaiju) were chos to determine the tissues' metagenomic profiles. Analysis for all software were performed with the default setting according to the instruction manuals provided by the developers.

Database building. DBs were doenloaded directly form the software’s homepage if a pre-built DBs was provided by the developers of the software (BLASTN, minikraken DB of Kraken2, Centrifuge, and Metaphlan3). DBs were build based on the standard instructions provided by the software’s manual if no pre-built DB were not released by software’s developers (CLARK, CLARK-s, Diamond, and Kaiju). DBs of some software were made available online with the contribution of the science community based on the instruction manual of the corresponding software. In this case, the DBs were also downloaded directly from the online resources (standard DB of Kraken2, maxikraken DB of Kraken2, and Bracken). Detailed information about DB building is available in Table I.

Custom Kraken2 DB Building ofKraken2’s custom DB was built following the instructions on building customized DB. All the libraries included in the standard DBS of Kraken2 (including NCBI RefSeq’s bacterial, archaeal, and viral libraries, along with the human genome and library of known vectors (UniVec\_Core)) were included in the DB with the addition of *R. norvegicus* genome (GCF\_015227675.2\_mRatBN7.2) and *R. rattus* genome (GCF\_011064425.1\_Rrattus\_CSIRO\_v1).

**Statistical analysis**. Metagenomic profiles were loaded into R for analysis using the package “phyloseq”. Pairwise significant difference assessments were performed by Wilcoxon signed-rank test implemented in R’s “rstatix” package, which is a non-parametric statistical hypothesis test used for comparing repeated measurements on a single sample (R Core Team, 2020). Alpha (Shannon, 1948; Simpson, 1949) and beta diversity (Bray and Curtis, 1957) indices (Whittaker, 1960) were used to describe the relationship of the microbes within and between samples, respectively, and were calculated with the R package "vegan" (Oksanen *et al.*, 2013). DA taxa analysis identify DA taxa between samples collected from two different tissues were assessed by the R package "DeSeq2" using the "Wald” test (Love, Huber and Anders, 2014), with reads classified under each species taxon normalized using the “poscounts” method. Data visualization for the metagenomics profiles were performed using the R package "ggplot2" (Ginestet, 2011).

**Results**

**Profiling compositions with different DBs**. To address the biases introduced from database selection during metagenomics profiling, four different databases (minikraken, standard, customized, and maxikraken) were used to classify the Rattus samples using Kraken2.

Three of the four databases was built previously and provided by the science community without charge (minikrakenV2, standard, maxikraken), while the customized databases were build following the protocols provided in the Kraken2 manual. Maxikraken2 DBs, although could be downloaded directly, requires over 150 GB memory the workstation used for analysis. While minikrakenV2, distributed by the developer of Kraken2, only requires less than 8GB. Customized database (60 GB) was built with the same composition of the standard database (53 GB), with the addition of the two Rattus genomes, which is the host species of the dataset. With 12 threads of CPU used on UGA’s high memory computing node, the building of the customized database took ~15 hrs (Table I.1) to complete the building process. Kraken2 will loaded database selected into the workstation for every analysis the software perform, thus the memory resources utilized during Kraken2’s analyses are directly correlated with the choice of the databases. Time of the analyses also changes with the selection of the databases, but the differences is only in the range of seconds.

The average number of total classified reads using these DBs range from 10,755 (SD: 20,651) using the minikraken DB to 21,402 (SD: 27,043) using the maxikraken DB (Table I.1). The numbers of reads classified under the the four highest taxnomy level (Domain), Eukaryota, Bacteria, Viruses and Archaea taxa, by each databases are presented in Figure 1a-d. The statistical significance of the differences in comparisons of the classification results for each Domain taxon classified by the four DBs were validated using the paired Wilicoxon signed rank test. The adjusted p value for all comparisons between DBs are available in Table I.2. Most of these pairwise comparisons between the results of DBs were significant. For example, the differences in the number of reads classified under the Bacteria taxon were significantly different for all pairwise DB comparisons (Figure 1b). For Eukaryota taxon, only the number of reads classified by the standard and customized DBs were found not significantly different in comparion (Figure 1a). Classification results for Viruses are more similar across DBs compare to that of Eukaryota and Bacteria, but still with 4 out of 6 comparison different significantly. In the end, for Archaea classification, only the classification results of minikraken were found significanly different when compared with the results of other DBs, the classification results of other three DBs were not different significantly between each other.

To understand how differences in classification results can directly impact the characterization of the microbial communities in each sample, we calculated two alpha indices (Shannon and Simpson) at the species level, characterizing the species richness (diversity) and the evenness of the microbial communities within each Rattus sample and compared them with the observed unique number of species identified by each DB (Observed). We have found that although the observed unique taxon were significanly different across the classification results of all four DBs (Figure 2a), only the Shannon index, which describes the species richness and evenness within a community, obtained frin minikraken DB were found significantly different when compared with the results of other DBs (Figure 2b, Table I.3). Morever, the Simpson index, which describes the evenness of the microbial communities within each sample, were also found mostly similarly between results of the four DBs. Only the evenness indices obtained from the results of the standard and customized DBs were found significantly different in comparison (Figure 2b, Table I.3).

In addition to the characterization of the microbial community within each sample, relationships between the microbial communities are also very important in metagenomics studies. The pairwise relationships between every two Rattus samples in the dataset were described with the Bray-Curtis index, and clustered hierarchically (Figure 3). The statistical significance of differences between the Bray-Curtis indices obtained from the results of different DBs were validated pairwisely using the paired Wilicoxon signed rank test (Table I.4). These indicies describing the between-samples relationships were found significantly different when using different DBs. Only the Bray-Curtis indices obtained from the results of maxikraken DB and customized DB were found not different significantly. Furthermore, hierarchical clustering of the samples, which describes the relationships between-samples compare to that of other samples in a dataset, has also impacted by the differences in profiling results using different DBs. We have observed that the three kidney samples (R22.K, R26.K, and R27.K) were found always clustering with one spleen sample (R26.S) in all four DBs’ classification, however, their relationships with spleen sample R27.S changes with the results of different DBs. For example, in the clustering with minikraken DB result, R27.S clustered more closely with spleen sample R22.S before clustering with the three kidney samples, but in the clustering results of other three DBs, R27.S clustered closely with the three kidney samples and R26.S while R22.S always clustered closely with sample R28.K. Despite the changes in the lower hierarchical levels, the two major clusters describing the general relationships between samples has not changed by using different DBs. Three Lung samples (R22.L, R26.L, and R27.L) has always clustered closely together away from other samples, while all Kidney and Spleen samples formed a separate cluster with R28.L.

**Profiling compositions using different metagenomics profiling software**.

The resources required to build database and to classify each sample diverges largely across software (Tabe I). Except for CLARK, CLARK-s, Diamond, and Kaiju, the analysis of the rest of the software could be ran with a pre-built database. With 12 threads of CPU used on UGA’s high memory computing node, building of CLARK’s database took over 42 hours to complete, with over 400 GB memory utilized (Table I). CLARK-s database, required to build on top of the CLARK’s database, took around 40 additional hour to complete, with around 300 GB memory utilized (Table I). Building of Diamond’s database, with the same computational setting, completed in ~2.4 hours utilizing ~ 8 GB, while Kaiju’s database took ~ 5 hours to complete utilizing ~115 GB of memory (Table I). As for analysis time, using 12 CPU on UGA’s high memory computing node, Diamond used ~5 hour on average to classify one sample and BLASTN used ~2 hr to classify one sample. Rest of the software could finish classifying one sample within a minute on average (Table I).

To compare the profiling compositions of the different sofware, we calculated the number of total classified reads for each sample and determined these samples’ profiling compositions at the domain, phylum, genus, and species levels. The average number of total reads classified by each software ranges from 10,955 using CLARK-s to 77,499 using Diamond (Table II.1). The number of unique taxa classified by each software also ranges from 18 taxa by Metaphlan3 to 4816 taxa by Kaiju (Table II.1). Furthermore, we have found that Metaphlan3 has not classified any reads in samples of Rattus R26 (R26.K, R26.L. and R26.S) and sample R22.L and R27.K, while other software has classified on average 1252 (SD: 1408), 32748 (SD: 32178), 133 (SD: 112), 111068 (SD: 113203), and 4011 (SD: 4325) reads with these five samples respectively (Table II.2).

The number of classified reads were break down at the domain level taxa, where we closely examined the number of reads classified into Eukaryota, Bacteria, Virsues, and Archaea by each software (Figure 4). The number of reads classified into the Eukaryota taxon has the largest differences across the classification results of different software, where only the number of reads classified by Centrifuge and Diamond were found not significantly different in this taxon. Furthermore, due to the limitation of their DB composition, Metaphlan3, CLARK-s, and Kaiju did not reported reads classified into the Eukaryota taxon. Compare to reads classified into the Eukaryota taxon, the number of reads classified into the Bacteria taxon were very similar across software. Only reads classified by CLARK and CLARK-s were found significantly different in Bacteria classifications with most other software (except for when compared with Metaphlan3 and Kaiju). The classifications of Viruses by different software, on the other hand, were divided into two groups, where each group are not significantly different within each other, but different with the results classified by the software in the other group. The first group includes the Virsues classification results of BLASTN, CLARK, CLARK-s, Metaphlan3, and Kaiju, and the second group includes the results of Kraken2, Bracken, and Centrifuge. Diamond classification didn’t identify any reads as Viruses in the Rattus samples. Archaea’s classification using different software are also very similar, only the classification results using Centrifuge were found significantly different with the classification results of most other software (BLASTN, Diamond, Kraken2, CLARK, and CLARK-s). In addition, Bracken and Metaphlan3 didn’t classify any reads into the Archaea taxon.

The read distribution at the Phylum and Genus level were also examined to increase the resolution of comparisons between software. At the Phylum level, the number of unique phyla taxa identified by each software ranges from 5 using Metaphlan3 to to 59 using Kaiju. We extracted the top 5 phylum taxa identified from each sample and combined reads classified to other phyla into the “p\_\_Other\_Phyla” (Figure S1). Top 5 Phyla has described a large percentage of read classification for all software’s classifications. However, the distribution of reads classified into different phyla taxa are different across software. For example, Virus taxon, “p\_Pisuviricota”, has contributed to over 85% (569/665) of the reads classified in sample R22.K using BLASTN, while this taxon was not identified by any other software’s classification. Nevertheless, Metaphlan3 has classified all of its reads in sample R22.K into “p\_\_Viruses\_unclassified”, and CLARK and CLARK-s has classified 63% (120/190) and 57% (95/166) of sample R22.K’s read to two different Virus taxa, “p\_\_Uroviricota” and “p\_\_Artverviricota”. Kaiju has also classified 21% of sample R22.K’s reads into “p\_\_Artverviricota” (34/157). Similar distributions in reads involving Virus taxa classification were also observed in sample R26.K, R26.S, and R27.K, where BLASTN classified 54% (657/1207), 20% (28/140), and 11% (422/3794) of reads into “p\_Pisuviricota”, respectively, CLARK and CLARK-s classified a large percentage of reads into Virus taxon “p\_\_Uroviricota” (CLARK: 71% (636/900), 31/76 (41%), and 18% (201/1099); CLARK-s: 18% (50/271), 18% (7/67), 10% (83/1334), respectively), but other software has only identified a small number or none reads into a Virus taxon (Kraken2 has classified 4 reads into taxon “p\_\_Uroviricota”). Except for differences in Virus taxa identification, the distribution of the Bacteria reads classified by BLASTn, Kraken2, Bracken, Centrifuge, CLARK, CLARK-s, and Kaiju are relatively consistent across samples without Virus taxa identified. The diversity of taxa identified by Metaphlan3 are significantly less than that of other software’s classification, only the most abundant taxa were captureing the majority of the classified reads with Metaphlan3 classification. For example, Metaphlan3 has identified 100% of sample R7.L’s reads as “p\_\_Proteobacteria, while other software has identified 29% (SD: 12%) of R27.L’s reads as “p\_\_Proteobacteria” on average with unique number of Phylum taxa identified range from 2 by Diamond (91% of reads classified as “p\_\_Tenericutes”) to 50 by Kaiju. Diamond’s classification is also showing differences in read classification when comparing with results of other software. The most notiable difference is the relative abudance of taxon “p\_\_Firmicutes” classified by Diamond across samples. In the lung samples, “p\_\_Firmicutes” was classified in 17% of R22.L (SD: 9%), 20% of R26.L (SD: 9%), and 14% of R27.L (SD: 8%), but Diamond has only classified 2% (133/4900) of reads as “p\_\_Firmicutes” in sample R26.L, while “p\_\_Firmicutes” taxon was not identified in R22.L and R27.L by Diamond. On the other hand, Diamond has identified a relative larger proportion of reads as “p\_\_Firmicutes” in samples R27.S (24%) and R28.L (19%) compare to that of most other software (R27.S: mean: 2%, SD: 2%; R28.L: mean: 3%, SD: 3%), except for the Centrifuge classification (R27.S: 24%, R28.L: 9%).

Moving down to the Species level classification, the number of reads classified under taxa (strains) with the same species was aggregated together to obtain the unique number of species classified by each software. Out of all software, metaphlan3 has classified the least number of species taxa with only 18 species. On the other hand, Kaiju has classified the most number of distinct Species taxa 4128 species (Table II.4). From the species level classifications, 9 species taxon were identified by all nine software (*Leptospira interrogans*, *Leptospira borgpetersenii*, *Faecalibacterium prausnitzii*,*Bordetella pseudohinzii*, *Bordetella bronchiseptica*, *Bordetella pertussis*, *Bacteroides uniformis*, *Phocaeicola vulgatus*, and *Bartonella elizabethae*). Centrifuge and Kaiju has the largest overlapping in the species taxa identified (2285 taxa), followed by Kraken2 vs. Centrifuge (1737 taxa) and vs. Kaiju (1723 taxa). The species-level classification of the three software has shared 1,379 species taxa in total. In addition, BLASTN has also shared 1253 species level taxa with Centrifuge, 1207 taxa with Kaiju, and 1126 taxa with Kraken2. CLARK and CLARK-s’s classification has also shared 1219 and 1059 species taxa wtith Kaiju specificially. To assess if different software has identified same species taxa as the most abundant taxa, species taxa with at least 10% of the reads from each sample were identified from each software’s classification. Metaphlan3 in this case, has identified most number of unique species taxa (18 taxa), while BLASTN and Kaiju has the least (7 taxa). CLARK vs. CLARK-s and Kraken vs. Bracken shared most number of taxa in this category (9 and 8 taxa, respectively). Two species taxa were identified by all software as the top ten percent most abundant species taxa which are *L. interrogans* and *Bartonella elizabethae*.

**Downstream analyses for microbial community characterization**

We have also obtained the Alpha and Beta diversities of the Rattus dataset at the species level to characterize the microbial communities each sample using different software’s classification results.

For Alpha diversities, which focuses on the within-species microbial communities characterization, we obtained three indices comparing each software’s characterization, 1) the observed number of unique species within each sample (Observed), 2) the Shannon index, which characterize the richness of each sample (Shannon), and 3) the Simpson index, which describes the evenness of the microbial abundance within each sample (Simpson) (Figure 5). The numbers of unique taxa observed from each sample were largely different using different sofwares. Out of the 36 pairwise comparison between different software, only 6 comparisons were not significantly different (Table II.4), which are BLASTN’s observed taxa with that of Kraken2, CLARK, and CLARK-s, comparison between CLARK and CLARK-s, and comparison between Centrifuge and Kaiju. Nevertheless, the Shannon index, obtained from these softwares are more similar than the observed numbers of taxa. 13 out of 36 comparison were found not significantly different. All the classifications of software found similar in observed taxa was also found not significant in difference for their Shannon indices. Shannon indices obtained with BLASTN’s classification was also found not different from the that of Bracken and Diamond. These software similar to the Shannon indices obtained from BLASTN was also found similar with each other, ex. Bracken vs. Diamond, Bracken vs. CLARK and CLARK-s, and Diamond vs. CLARK and CLARK-s, and etc. For the evennnes within each sample, the Simpson’s index were least impacted by the differences in classification results across software. Only 7 out of 36 comparison were found significantly different in Simpson indices. Most of these significantly different comparisons were identified between CLARK-s (3/7) and Centrifuge (4/7) with other softwares or between each other.

In addition to within-sample characterization, the pairwise between-sample relationships were measured by the Bray-Curtis indices and clustered using the hierarchical clustering method. Validating comparison using the paired Wilicoxon signed rank test, we identified that the pairwise between-sample relationships evaluated using BLASTN was not different from that evaluated with Kraken2, Bracken, and Centrifuge, and between-sample relationships evaluated using CLARK and CLARK-s are not different from most other softwares except for BLASTN, Centrifuge, and Metaphlan3, separating these software into two groups. Metaphlan3, with 5 out of 12 samples unclassified completely, was significantly different in the between-sample relationships with that obtained from other software (Table II.5). We further explored the relationships between-samples compare to that of other samples using using hierarchical clustering (Figure 6). Except for the clustering using the Metaphlan3 classification, the classification of rest software has clustered the Rattus samples into two large clusters, first cluster included three lung samples (R22.L, R26.L and R27.L) and second cluster with all the Kidney and Spleen samples as well as the Lung sample of Rattus subject R28 (R28.L). However, the smaller clusters formed inside the second cluster varies among software. For example, BLASTN’s classification has clustered three Kidney samples (R22.K, R26.K, and R27.K) closely with each other before clustered further with other samples. This cluster was only observed in Centrifuge’s clustering. While in the clustering of other software, these three kidney samples were always clustered closer with other samples before clustered with each other. For example, in Diamond, CLARK, and CLARK-s, one of the three kidney samples has always clustered with R26.S before clustering with each other. In classification of Kraken2, Bracken, and Kaiju, both R26.S and R27.S was clustered with one of the kidney samples before the kidney samples clustered with each other. However, in BLASTN’s classification, both R26.S and R27.S was clustered with the rest of the Spleen sample first (R22.S and R28.S) before clustered together with the three Kidney samples.

**DA taxa identification**

We have also identified the DA taxa between the samples of different tissues. The microbial communities of the lung samples were found most distinctive from that of spleen and kidney samples despite the differences in the classification results reported by different software. Therefore, we have started with identify the DA taxa between the lung samples and the kidney samples.

Since Metaphlan3 wan not able to classify 2 lung samples and 2 kidney samples, we will exclude the classification analyses of Metaphlan3 from this analysis. At the species level, the number of DA taxon identified using the classification results of different software ranges from 10 (Diamond) to 596 (Centrifuge), with more taxa significantly higher in abundance in the Kidney samples than that of Lung samples with all software’s classifications (Figure S2). Five significantly abundant species was found shared by the classification results of all software (*Bordetella pseudohinzii*, *Bordetella bronchiseptica*, *Leptospira interrogans*, *Leptospira borgpeterseni*, and *Mycoplasm pulmonis*). Kaiju, Centrifuge, BLASTN have the most number of DA taxa mostly distinct to themselves (390, 376, and 56 taxa, respectively) (Figure 7a). Furthermore, although Centrifuge identified the largest number of DA species taxon, Kaiju has identified the most number of unique phylum taxa (42), which means many of Centrifuge’s DA species has the same phylum taxonomy taxa (Figure 7a). To obtain a more generalized overview for the DA taxa identified from the classification of each software, we aggregated the species taxa into the phylum level and visualized the presence and absence of each phylum taxon as the DA taxa across different software in Figure 8. At the Phylum level analysis, taxa “p\_\_Spirochaetes”, “p\_\_Bacterodietes”,“p\_\_Protebacteria”, and “p\_\_Tenericutes” was found present in the results of all software. Diamond was missing four taxa that were identified by rest of the software ("p\_\_Aquificae”, “p\_\_Fusobacteria”, “p\_\_Firmicutes”, and “p\_\_Cyanobacteria”). Kaiju and Centrifuge were th only two software reported virus taxa as DA. Both software reported virus taxon “p\_\_Negarnaviricota”, and Kaiju reported “p\_\_Nucleocytoviricota” and “p\_\_Uroviricota”, distinctively. Archaea taxa was only reported by Kaiju, Centrifuge, and BLASTN’s . All three software have reported "p\_\_Euryarchaeota”, and both Kaiju and Centrifuge reported “p\_\_Candidatus Micrarchaeota” and "p\_\_Candidatus Lokiarchaeota”. Finally, Kaiju uniquely “p\_\_Candidatus Thermoplasmatota”.

The DA taxa identified between lung samples and spleen samples were similar with those identified in the between lung samples and kidney samples. Kaiju in this case has identified the most number of DA species (484 taxa), while Diamond has identified least (44 taxa). All of the DA taxa were more abundant in the lung samples. Six species were overlapping between the DA taxa identified by the classifications of all software (*Mycoplasm pulmonis*, *Mycoplasma bovoculi*, *Mycoplasma neurolyticum*, *Bordetella pseudohinzii*, *Bordetella bronchiseptica*, and *Bacteroides uniformis*), three of the overlapping species were also identified as DA species overlapped among all software during Lung vs. Kidney samples comparison. Kaiju still has the most number of distinct DA species taxa (335 taxa), followed by centrifuge (268 taxa), and BLASTN (46 taxa) (Figure S10). On the Phylum level, “p\_\_Bacterodietes”, “p\_\_Tenericutes”, “p\_\_Cyanobacteria” ,“p\_\_Protebacteria”, and “p\_\_Firmicutes” was identified by all software as DA. Taxa "p\_\_Aquificae”, "p\_\_Actinobacteria”,and “p\_\_Fusobacteria” were identified in by all software except for Diamond. Archaea phylum, "p\_\_Euryarchaeota”, was still the Archaea taxon identified by BLASTN, Centrifuge, and Kaiju, rest of the Archaea taxa were either only identified by Kaiju and Centrifuge, or Kaiju alone. Virus taxon, “p\_\_Negarnaviricota”, was only identified by Centrifuge as differentially abundant, while Kaiju identified virus taxa only reported “p\_\_Nucleocytoviricota” and “p\_\_Uroviricota”. Morever, in this comparison, CLARK has also reported virus taxon, “p\_\_Uroviricota”, as significantly abundant.

Finally, we futher identified the DA species between Kidney and Spleen samples. The number of species identified ranges from 6 by Diamond and 57 by BLASTN (Table II.8). More taxa was identified significantly abundant in the Kidney samples compare to the Spleen samples, especially in the genus level (Figure S8). Kaiju, the software identified the second highest number of distinct DA taxa at the species level, has five out of ten distinct taxa reported as Viruese (Figure S5). In general, Only 1 species (*Leptospira interrogans*) and 4 phylum taxa (“p\_\_Spirochaetes”, “p\_\_Bacteroidetes", "p\_\_Cyanobacteria”, and “p\_\_Proteobacteria”) was found overlapping with all software’s classifications (Figure S5, Figure S6). The Phylum taxon “p\_\_Firmicutes” was identified from the classifications of all software as the DA taxon except for Diamond. Kaiju has identified the only virus taxon, “p\_\_Negarnaviricota”, as a DA taxon.

***Leptospira* detection.**

With the use of the nine software, Leptospira was identified in the three tissues of all four subjects, but each software has reported Leptospira in different samples (Table I). Centrifuge is the only software reported Leptospira in all 12 Rattus sample, where 8 unique Leptospira species has been identified (8 from the pathogenic group, 1 from the saprophytic group). Kaiju has also identified Leptospira from 9 out of 12 samples with 8 unique species (7 from the pathogenic group, 1 from the saprophytic group). Kraken2, following Centrifuge and Kaiju, has classified 6 Leptospira in 6 samples with 3 unique species all from the pathogenic group. Except for Metaphlan3, all software has identified Leptospira from R22.K and R28.K, which has 31 (SD: 3) and 84,344 (SD: 2.2) reads classified under Leptospira on average, respectively. BLASTN and CLARK has also identified Leptospira from R22.L, which was also identified by Centrifuge, Kaiju, and Kraken2. Metaphlan3 has only identified Leptospira in R28.K. All samples identified by at least three software has at least 30 reads classified under Leptospira in total (Table SIII.3). Samples that were only identified by Kaiju or Centrifuge has only 2 (R27.K, SD: 1) to 15 (R26.L, SD: 2) reads classified under Leptospira on average. In addition to differences in Leptospira diagnosis caused by the use of different software, diagnosis of Leptospira was different when different databases were used for Kraken2’s classification (Table SIII). Kraken2’s analyses with the maxikraken DB has identified Leptospira in all samples, while standard and customized DB has identified Leptospira in two Lung samples (R22.L and R27.L). Standard DB has also identified Leptospira in the three Spleen samples (R22.S, R27.S and R28.S). In addition to the metagenomics approaches, the diagnosis of Leptospira in the kidney samples was validated using three traditional methods (PCR/DFA/Culture), Leptopsira was identified in samples R22.K and R28K by all three methods, but only identified by PCR in samples R26.K.

**Discussion**

Profiling the microbial taxonomies from biological specimens allows a better understanding of the microbial communities of the samples collected for different fields of studies (Coyte, Schluter and Foster, 2015; Gilbert and Lynch, 2019). The field of metagenomics, developed with the advancement of NGS technologies, allows scientists to build a complete and discriminatory microbial profile with virus, archaea, and bacteria taxa for samples collected from their target of interest (Jovel *et al.*, 2016). These metagenomic profiles can be used to detect relevant pathogens in clinical and epidemiological investigations (Qin *et al.*, 2012; Knights, Lassen and Xavier, 2013) and to observe the interactions between a micro-ecosystem and its changing environment in the ecological contexts (Handley, 2019).

In this study, microbial profiles of twelve samples collected from 4 wild rat subjects were classified. These rats were captured in the Caribbean island of St.Kitts, and are the major reservoirs contributing to the transmission of the pathogenic *Leptospira* on the Leptospirosis endemic island (Boey *et al.*, 2019). Rats harbor this bacteria in their kidneys are a significant source of environmental contaminations (Saito *et al.*, 2013; Rawlins *et al.*, 2014; Costa *et al.*, 2015; Boey, Shiokawa and Rajeev, 2019; Rajeev *et al.*, 2020). In addition to kidney samples, we also classified the microbial profiles of samples from rat lungs and spleen. We identified the microbial profiles of these samples, which contain many potential rodent pathogen sequences, using nine different shotgun metagenomics sequencing taxonomic classification software. The differences in the classification outputs were compared and analyzed to address how the use of different taxonomical profiling software on the same dataset could lead to diverged diagnosis in *Leptospira* pathogen, and also affect the results of microbiome characterization, which lead to different biological conclusions in the downstream analyses.

Previous benchmarking studies (Escobar-Zepeda *et al.*, 2018; Ye *et al.*, 2019a) have performed comprehensive analyses on these software’s speed and performances (sensitivity, specificity, precision, and accuracy). However, these benchmarks have been generally based on *in silico* datasets or with the support of laboratory synthetic samples. Differences among the microbial taxonomical profiles classified with different tools can be the result of reporting false positive or false negative taxonomies using different algorithms, DBs, or software. These differences sometimes seem negligible in the benchmarking studies for tools with similar algorithms but can lead diverging biological conclusions in the downstream analyses depending on the questions being asked. These biases originated from these differences have been understudied; therefore, it is crucial to demonstrate these biases with real biological data, to raise awareness for their existences and to identify the potential factors that lead to the incorrect biological conclusions in a metagenomics study.

**Biases Introduced by DB Selection**.

Incorrect taxonomical profiling of the collected samples’ microbial community can start from choosing an incorrect DB. All current taxonomical profiling software requires a large number of computational resources for DB building and storage. Some software, such as Kraken2, provide an alternative pre-built DB for users with inefficient computing resources, which minimize the size of the DB to be loaded into a machine with RAM as low as 8 GB. This DB is built with all libraries included within the standard Kraken2 DBs but down samples the size of the sequences included using a hash function. There were also multiple versions of Kraken2’s DBs provided by the science community that can be easily downloaded and updated frequently. For example, the Langmead lab builds the most recent version of Kraken2’s standard database based on NCBI’s RefSeq library routinely. In addition, the Loman lab has built a Kraken2 DB with the inclusion of the draft genomes that were not included in the Refseq library. Both of these two Kraken2 database were available freely to use online, and replacing the workload of building a database from scratch. However, all three databases mentioned above has included human genome as the only Eukaryotic genome in the database, which are not the host of our dataset. The biases introduced from host genomes included in the database for metagenomics analysis has been address previously (Pereira-Marques *et al.*, 2019). Therefore, we have built a separate database with the inclusion of the two Rattus hosts genomes on top of the standard database as the customized database for our dataset. We found that although the number of reads classified using different databases different significantly, the characterization of each sample’s microbial communities won’t be largely biased by the use different databases. In our analyses, only the richnesses of the samples (Shannon indices) obtained from the miniKraken DB were different significantly with that of other software. The evenness of the samples (Simpson indices) were mostly consistent across classifications of different DBs. Although the microbial communities characterized within-sample was not biased by the use of different DBs, the relationship of the microbial communities between-samples were statistically different across the classification results of most databases. By hierarchical clustering samples based on their pairwise relationships, we found that only the higher level clusters descrbing the most distinctive relationships between samples were consistent across the classifications of all DBs. Sophiscated relationships between samples were altered by the biases introduced from DB selection.

**Resources Required to Use Different Software**

The metagenomics software can be classified into two different categories, alignment-based and alignment-free. The first category, represented by BLASTN and Diamond in our analysis, are the most traditional methods available for metagenomics profiling analyses. Although were found high in sensitivity, this type of software requires extremely large computational resources and are really slow in analysis. The second category of software, which were alignment-free, were developed in response to the deficiency of alignment-based software to decrease the computational resources and time utilized for analysis.In our analysis, BLASTN and Diamond, were the two most time expensive software out of all software. These two software took ~2 hours and ~5 hours on average to complete the analysis for one sample, while other software took less than a minutes for doing the same task. However, the time and resources required to build the DBs of the alignment-free software became the trade-off for the speed of the analysis itself. For example, the building of CLARK’s database took almost 2 days with 400 GBs of memory used. Forunately, most of the software included in our study has pre-built DBs distributed with the release of the software (except for CLARK, CLARK-s, Diamond, and Kaiju). However, if the analysis requires the identification of taxa that are not included in these pre-builit DBs, the time and resources added to the metagenomics profiling analysis will increase significantly.

To compare each metagenomics profiles classified by each software, we chose the standard DBs provided by the developers of these software. If the standard DBs was not indicated, we built the DBs with the genomes of Bacteria, Archaea, Viruses, and Human available in NCBI’s RefSeq library, which is the compositions of the databases for most of the pre-built DBs. CLARK-s’ DB was required to built on top of a CLARK DB of the same composition, but when the CLARK-s’ DB was intended to build on top of the CLARK DB with the genomes of Bacteria, Archaea, Viruses, and Human, the building was suspended by the software with the error message “the number of targets exceeds the limit (16383)”. This limitation was reported to the developer of CLARK-s, but was has not been resolved yet by the time this manuscript was drafted. We bypassed the limitation by building the DB with Bacteria, Archaea, and Virsues genomes separately, and combine the classifications using each DBs at end of the analysis. In addition, Metaphlan3, which identifies the microbial taxon with marker genes, does not have an option to build a customized DBs, only the marker DB distributed by the developer could be used for profiling.

**Biases in Micorbial Profiles Introduced from Software Selection**

At the Domain level, Eukaryota taxon has contributed most into the differences between the classifications of different software. Almost all pairwise comparisons between the Eukaryota profiles of two software were found significantly different between each other. Compare to reads classified under Eukaryota, the number of reads classified under Bacteria, Viruses and Archaea taxa by different software were much more similar between software. The classifications of Centrifuge, CLARK, and CLARK-s were frequently identified significantly different from other softwares in comparisons for number of reads mapped to Bacteria and Archaea. The classifications of Viruses, on the other hand, were only found separated into two groups, where the classifications of software within a group were similar (group1: BLASTN, CLARK, CLARK-s, Metaphlan3 and Kaiju; group2: Kraken2, Bracken and Centrifuge). Diamond didn’t identify any reads as viruses from the Rattus dataset.

This division in virus classifications was further validated by the virus classifications at the lower taxonomy levels. The samples with reads classified under virus taxa by group1 software were not profiled by software in group2. Although software in group1 were more sensitive in virus identification than that of group2 software, the exact virus taxa and the number of reads classified under each virues taxa using different group1 software were not consistent. The virus taxon identified by BLASTN in high abundance was not identified by any other software included in the analysis. Except for the samples with virus classifications, the profiling of bacteria taxa were found mostly consistent across the software at both Phylum and Genus level. Only the classifications of Metaphlan3, which could only identify a few taxa from each sample with high abundance, and Diamond, which reported conflicting profiles in Firmicutes identification at the Phylum level (Bacillus at Genus level) with the classification of all the other software, were different by observation.

Compare to phylum and Genus level, the classifications at the Species level were more diverged across software. Although most software has reported more than 1,000 unique Species taxa from the Rattus profiles (except for Bracken and Metaphlan3), only 9 species were found identified by all software included in this analysis, and only 2 species were found overlapped in the taxa with at least 10% in relative abundance identified from the classification of each software.

**Microbial Community Characterization**

Although the differences in microbial profiles classified by different software were described extensively at different taxonomical level, we would like to know how does these differences could impact the characterization of each sample’s microbial community quantitively. We first characterized the richness and evenness of each sample’s within-sample microbial diversity at the species level. The differences across the richness of each samples’ microbial community were significant in the majority of the comparisons between software. Most of the significant different comparisons were found between the classifications of Kraken2, Metaphlan3, Centrifuge, and Kaiju with other software. The evenness of the samples, on the other hand, were mostly not affected by use of different software.

As for the relationships between samples in the dataset, the indices describing the relationships between every two samples were found significantly different in most of software’s comparisons. However, the most distinctive difference between the lung samples and other tissue samples were captured by most software (except for Metaphlan3), but the descriptions of the more subtle relationships between samples were not reported consistently across software.

**Differences in Differential Abundant Taxa**

In order to address potential biases introduced from software selection with biological significance, we had identified the DA taxa between samples of different tissues in a pairwise fashion. From the between-sample relationship analyses, all software has reported that the micobial communities of lung samples were distinct from that of kidney and spleen. Following this obersevation, we wanted to know what are the taxa that contributed most to the differences in the microbial profiles between different rat tissues? Were DA taxa identified from lung vs. kidney and lung vs. spleen samples similar to each other? Were the number of DA taxa reported from kidney vs. lung samples comparison less than those reported when compared to the lung samples? Metaphlan3 were excluded from this analyses due to not classifiying 5 out of 12 samples in the dataset. The classifications of all DA taxa reported at the species level were largely different when the profiles of different software were used. The largest range in the number of differentially abundant taxa reported by different software were found in the anlysis between lung and kidney samples, where the software with least DA taxa identified, Diamond, reported 10, and the software identified most DA taxa, Centrifuge, reported 596. Despite the large differences in the number of taxa identified, there were still small number overlapping species identified across the results of all software. We have also found similarities in the software-overlapped DA taxa between lung vs kidney and lung vs. spleen analyses, where two *Bordetella* species and a *Mycoplasm* species were overlappingly reported by all software in both analyses. More DA identified were overlapped across software at the Phylum level. In addition to the overlapped DA taxa, Kaiju and Centrifuge were more likely to report more taxa as differentially abundant compare to the analyses of other software. These two software were also the only two software reported both viruses and archaea taxa as DA taxa (BLASTN only reported Archaea, and CLARK only reported virueses). Diamond was found least sensitive in differentially abundant analyses for all three comparisons between tissue samples, where phylum taxa identified by all the other software were frequently not identified by Diamond.

Leptospira Diagnostic Sensitivity Comparison

To assess the sensitivity of shotgun metagenomics as a tool for pathogen diagnosis, we identified the presence of the zoonotic pathogen *Leptospira* in all of our tissue samples. Centrifuge and Kaiju were found most sensitive in diagnoising Leptospira, where Centrifuge has reported the presence of Leptospria in all 12 samples. Except for Centrifuge and Kaiju, rest of the software were consistent in Leptospria identification in only two of the kidney samples (R22.K and R28.K) and one Lung sample (R22.L), where the reads classified under Leptospira were relatively more abundant. Since Leptospira pathogens were mainly deposit in the kidney of rats before infecting or contaminating other mammals or environment through urination (Adler and de la Peña Moctezuma, 2015), we diagnosed the presence of *Leptospira* using three traditional methodologies (PCR/DFA/Culture) in the kidney samples alone i. Through traditional method, Leptospira was identified in sample R22.K, R27.K, and R28.K, in which R27.K was only identified by PCR. These results suggest that most software included in our analysis has similar sensitivity in Leptospira identification with traditional methods, except for PCR. In addition, Centrifuge has reported the presence of Leptospira identified in sample R26.K, which was not identified by any other software or a traditional method. This identification could be due to Centrifuge’s better performances in sensitivity, or as a result of false positive reporting. Furthermore, we found Leptospira was also identified in sample R26.K if maxikraken DB was used for Kraken2’s analyses. Kraken2 with maxikraken DB has also reported *Leptospira*’s presence in all 12 samples. We hypothesize that sensitivity of Leptospira’s diagnosis may improve with more draft genomes of Leptospira included in the database, because most of Leptospira species’s genomes were only available in the draft format.

The inconsistencies found between the results of different metagenomic software show that significant biological conclusions from metagenomic profiling analyses have the potential to be only the artifacts of the software’ algorithms. Shotgun metagenomics sequences might be too short for current taxonomical profiling software to differentiate microbial taxonomies between similar genomes (Tran and Phan, 2020). The use of real-world datasets has the advantage of addressing this challenge in metagenomic studies from the users’ perspective, reminding the investigators to stay skeptical with the classification results obtained from the profiling software. On the other hand, benchmarking the software’ performances with the real-world dataset, in contrast to using *in silic*o datasets, has the limitation of lacking knowledge about the true microbial compositions within each sample, which means we could not evaluate the performance of software based on their degrees of accuracy and sensitivity nor giving direct suggestions on software’ selection. In addition, metagenomics profiling has been broadly utilized in many fields of studies, including clinical, pharmaceutical, as well as ecological studies. Each field utilizes microbial profiles differently based on the biological question proposed. Our choice of the real-world dataset could only address a limited number of software selection biases. We suggest researchers from different study fields to be aware of the possible error-prone conclusions made from metagenomics profiling analysis, and evaluate it objectively comparing it to other traditional methods (e.g. PCR, culture, or serotyping).

Advancement in sequencing as well as computational technologies allows modern-day biological research to move to a brand-new era. However, while benefiting from the powerfulness and convenience of technologies, we should always critically analyze and validate software outputs based on our prior knowledge.

**Acknowledgements**

This work was supported by the National Science Foundation under Grant No. DGE-1545433 to R.X. and startup funds to L.C.M.S. from the University of Georgia Office of Research. The sample collection, sequencing and analysis was done during S.R.’s tenure at the Ross University School of Veterinary Medicine, Saint Kitts and it was supported by internal grants from the Center for One Health and Tropical Medicine. We also would like to thank Dr. Kanae Shiokawa for her help with collection and processing of rat specimens.

**Conflicts of interest**

No conflict of interest declared.

**Repositories**

The raw sequence files (FASTQ) were submitted to the NCBI Sequence Read Archive under the Bioproject accession number: PRJNA717669. The individual isolates can be accessed under the following Biosample accession numbers: SAMN18507082 - SAMN18507091. All scripts for this publication are freely available on the following Github link: <https://github.com/salvadorlab/MetagenomicsToolsEvaluation>.

**Data summary**

The raw sequence files (FASTQ) were submitted to the NCBI Sequence Read Archive under the Bioproject accession number: PRJNA717669. The individual isolates can be accessed under the following Biosample accession numbers: SAMN18507082 - SAMN18507091. The short-read archive accession numbers are listed in Table S1.

**Ethical Approval**

Rats were captured following protocols approved by the Ross University School of Veterinary Medicine (RUSVM) IACUC (approval # 17-01-04).

**Supporting Information**

Supporting\_document.docx

TableS1\_sample\_data\_information.xlsx

TableS2\_Kraken2\_db\_domain.xlsx

TableS3\_Kraken2\_db\_comparison.xlsx

TableS4\_Kraken2\_std\_vs\_cus\_genus.xlsx

TableS5\_software\_domain\_phylum\_readsSummary.xlsx

TableS6\_software\_genus\_readsSummary.xlsx

TableS7\_Alpha\_index\_values.xlsx

TableS8\_AlphaIndexComparison.xlsx

TableS9\_Beta\_index\_values.xlsx

TableS10\_betaIndexComparison.xlsx

TableS11\_sigTaxa\_foldChange\_pvalue.xlsx

**References**

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

**Tables**

**Table 1.** *Leptospira* detection from kidney samples using Kraken2, CLARK, CLARK-s metagenomic profiling and traditional laboratory techniques: Polymerase Chain Reaction (PCR), Direct Fluorence Antibody (DFA) test, and culture. The relative cut-off for *Leptospira* diagnostic in metagenomic samples are 0.1%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample ID | Kraken2 | CLARK | CLARK-s | PCR/DFA/Culture |
|  | Absolute (Relative %) No. Reads | | |  |
| 22K | 53 (5.60%)\* | 31 (0.03%) | 31 (0.15%)\* | +/+/+ \* |
| 26K | 0 (0%) | 0 (0%) | 0 (0%) | -/-/- |
| 27K | 0 (0%) | 0 (0%) | 0 (0%) | +/-/- \* |
| 28K | 117,315 (96.89%)\* | 72,255 (90.53%)\* | 72,211(96.27%)\* | +/+/+ \* |

\**Leptospira* presence in the sample

**Figures Legends**

Figure 1. Domain level taxonomy profiles in samples from *Rattus rattus* and *Rattus norvegicus* tissues using different Kraken2 DBs. Each row of panels represents the absolute and correspondent percentage of compositional profiles for ‘Eukaryota’(), ‘Bacteria’ (), ‘Viruses’ ()\*, ‘Archaea’(), ‘Unclassified’ (), and ‘Other Sequences’() for each sample tissue (K: kidney, L: lung, S: spleen) from *Rattus norvegicus* (R22, R26, R27) and *Rattus rattus* (R28) using prebuilt MiniKraken2\_v1\_8GB (A, B), standard Kraken2 (C, D), and customized Kraken2 (E, F) databases. Each column in every sub-figure from left to right represents the domain level microbial compositions for samples R22\_K, R26\_K, R27\_K, R28\_K, R22\_L, R26\_L, R27\_L, R28\_L, R22\_S, R26\_S, R27\_S, R28\_S. The sum of all read classified in each sample’s compositional profile is corresponding to 100%. \*Even though viruses were classified by an independent taxonomic system, in this context, we will treat them as if they were an independent domain.

**Figure 2. Kraken2 DBs statistical comparison.** The number of reads of each tissue sample (K: kidney, L: lung, S: spleen) from *Rattus norvegicus* (R22, R26, and R27) and *Rattus rattus* (R28), which were classified into “Unclassified” (A), “Other Sequences” (B), and into four domains (Eukaryota (C), Bacteria (D), Viruses (E), and Archaea (F)) by three different Kraken2 databases (MiniKraken2\_v1\_8GB, Standard, and Customized). The number of reads classified into each category for each sample are presented in Table S2. All pairwise statistical comparisons within this figure were performed with a Wilcoxon signed-rank test. Samples: R22\_K (), R22\_L (), R22\_S (), R26\_K (), R26\_L (), R26\_S (), R27\_K (), R27\_L (), R27\_S (), R28\_K (), R28\_L (), R28\_S (). \*\* p-value < 0.01, \*\*\* p-value < 0.001

**Figure 3. Phylum level microbial composition profiling in samples from *Rattus rattus* and *Rattus norvegicus* using Kraken2, Clark, and CLARK-s.** Each row of panels represents the absolute and correspondent percentage of microbial compositional profiles at the phylum taxa under ‘Bacteria’, ‘Viruses’\*, and ‘Archaea’ classified by the three different software Kraken2 (A, B), Clark (C, D), and CLARK-s (E, F)). The sum of all read classified in each sample’s microbial compositional profile is corresponding to 100%. Each column in every sub-figure from left to right represents the domain level microbial compositions for sample R22\_K, R26\_K, R27\_K, R28\_K, R22\_L, R26\_L, R27\_L, R28\_L, R22\_S, R26\_S, R27\_S, and R28\_S. Phylums present in the figure: Proteobacteria (), Actinobacteria (), Cyanobacteria (), Firmicutes (), Bacteroidetes (), Uroviricota (), Tenericutes (), Spirochaetes (), Fusobacteria (), Chlamydiae (), Aquificae (), Chloroflexi (), Thermotogae (), Artverviricota (), Peploviricota (), Deinococcus- Thermus (), Planctomycetes (), Verrucomicrobia (), Nitrospirae (), Thaumarchaeota (), Crenarchaeota (), Euryarchaeota (), Nucleocytoviricota (),Cossaviricota (), Gemmatimonadetes (), Acidobacteria (), Candidatus Gracilibacteria (), Dictyoglomi ().\* For visualization purpose, only phylum taxa with percentage >=0.1% are included.

**Figure 4. Alpha Diversityacross *Rattus* tissues.** The alpha diversity of each sample’s microbial composition is described by three indices (Shannon (A), Simpson (B), Inverse Simpson (B)), which characterize each sample’s microbial composition based on richness, evenness, and species dominance, respectively. All three indices were calculated based on the absolute number of microbial reads (including all reads classified under the genus taxa of ‘Bacteria’, ‘Viruses’, and ‘Archaea’) classified by Kraken2, CLARK, and CLARK-s. All pairwise statistical comparisons within this figure were performed with a Wilcoxon signed-rank test. Samples: R22\_K (), R22\_L (), R22\_S (), R26\_K (), R26\_L (), R26\_S (), R27\_K (), R27\_L (), R27\_S (), R28\_K (), R28\_L (), R28\_S (). \*\* p-value < 0.01, \*\*\* p-value < 0.001.

**Figure 5. Between Sample Microbial Composition Dissimilarity.** Bray-Curtis indices quantify the dissimilarity between two samples’ microbial compositions. Higher Bray-Curtis values indicate a high level of dissimilarity between the two samples’ microbial composition. They were calculated based on the absolute number of reads classified under the genus taxa of Bacteria, Viruses, and Archaea from microbial compositions of Kraken2 (A), CLARK (B), and CLARK-s (C). The red and yellow colors show low and high levels of dissimilarity, respectively. Hierarchical clustering was used to cluster together samples that have similar microbial compositions (dendrograms on the left and top of the heatmaps). Pairwise statistical comparisons between Bray-Curtis indices were calculated from each software’s microbial compositions using Wilcoxon signed-rank tests. \*\* p-value < 0.01, \*\*\*\* p-value < 0.0001.

Figure 6. Sample Clustering with Multidimensional Scaling (MDS) analyses. Bray-Curtis indices were calculated from the microbial composition classified by Kraken2 (A), CLARK (B), and CLARK-s (C) and visualized using a dimension reduction method, MDS. MDS clusters samples with similar microbial composition based on Bray-Curtis indices’ pairwise distance matrix. Samples: R22\_K (), R22\_L (), R22\_S (), R26\_K (), R26\_L (), R26\_S (), R27\_K (), R27\_L (), R27\_S (), R28\_K (), R28\_L (), R28\_S ( ).

Figure 7. Differentially abundant phylum and genus taxa. Genus taxa found differentially abundant in *Rattus* tissue, kidney (R22\_K, R26\_K, and R27\_K, R28\_K), and lung (R22\_L, R26\_L, R27\_L, and R28\_L), by Kraken2 (A), CLARK (B), and CLARK-s (C) were found above along with their log-fold change in abundance across these two tissues. Each circle corresponds to a genus taxa that were found significantly different in abundance across kidney and lung samples’ microbial compositions, and the circle color represents the corresponding phylum. Phylum: Spirochaetes(), Proteobacteria(), Tenericutes(), Firmicutes(), Bacteroidetes(), Cyanobacteria(), Peploviricota().