# Higher classification sensitivity of short metagenomic reads with CLARK-*S*[*]

# Supplementary Material

Rachid Ounit and Stefano Lonardi

Department of Computer Science & Engineering,
University of California, Riverside,
900 University Ave,
Riverside CA 92521, USA

[*] A preliminary version of this work was presented at the *Workshop on Algorithms in Bioinformatics* (WABI) in Atlanta, GA, 2015, and included in its *Proceedings*.

## Supplementary Note 1: Generation of synthetic datasets and negative controls

In this note, we describe how we created the synthetic datasets used for the evaluation of the three tools we tested. To produce synthetic reads we have considered the organisms that were reported present by different published studies in real microbial habitats. We consider the habitats related to mouth, city parks/medians, gut, indoor, and soil (listed below).

- **"Buc12"**: As reported in [4,6], the dominant genus found in the oral cavity is *Streptococcus*. Study [4] also reports the presence of the *Haemophilus influenzae, Haemophilus parainfluenzae, Neisseria subflava* and *Veillonella dispar*. Thus, we chose these four species along with eight species selected from the *Streptococcus* genus (see Supplementary Figure 1).

- **"CParMed48"**: Forty-eight species were selected from *Proteobacteria, Acidobacteria, Bacteroides, Actinobacteria,* and *Planctomycetes.* These are the dominant phyla reported in [11] in city parks and medians in Manhattan (see Supplementary Figure 2).

- **"Gut20"**: This dataset contains the twenty species described in the Supplementary Table 1 of [8] (see Supplementary Figure 3).

- **"Hous31"**: Bacteria typically found indoor are *Streptococcaceae, Lactobacillaceae,* and *Pseudomonadaceae* (due to human activities), and also *Intrasporangiaceae* and *Rhodobacteraceae* (due to the environment), as reported in [12] (see Supplementary Figure 4). We selected thirty-one species from these microbial families.

- **"Hous21"**: We selected twenty-one species from the dominant organisms reported in [1] found in the bathroom and kitchen, namely *Propionibacterium acnes, Corynebacterium, Streptococcus,* and *Acinetobacter* (see Supplementary Figure 5).

- **"Soi50"**: We selected fifty species from the dominant genera reported in [3], namely *Acidobacteria, Actinobacteria, Bacteroides, Proteobacteria* and *Verrucomicrobia* (see Supplementary Figure 6).

A seventh dataset **"simBA-525"** containing reads randomly selected from 525 bacterial/archaeal species was also added (see Supplementary Figure 7). All the supplementary figures have been generated using Krona [9].

Datasets generation:
We obtained reference genomes (whole genome sequences) from the full NCBI/RefSeq database (~650 billion of nucleotides, containing more than 57,000 genomes distributed in 14,675 species, downloaded on February 9, 2016), then we used the ART read simulator [6] to create synthetic reads from the list of species listed above. We ran ART with default quality base profile and error parameters, length 100bp, and coverage 30x. These seven datasets represent a total of 647 species (see Supplementary Table 1 for statistics on these datasets).

Datasets of unambiguously mapped reads:
To create datasets of unambiguously mapped ~~read filtered variants (i.e., datasets without ambiguously mapped reads)~~ for each of these seven datasets, we used the method described in Supplementary Note 2.

Negative control samples:
To generate negative controls, we created three datasets (named "LM", "MH1", "MH2") composed of reads that do not exist in any genomes in the NCBI/RefSeq database (see Supplementary Table 1). To build these datasets, observe that if a DNA fragment of 100 bps contains at least one *k*-mer that does not appear in any genomes in the full NCBI/RefSeq database then it does not exist in any of these genomes. In other words, if each read contains one *unassigned k*-mer for the full NCBI/RefSeq database then the read does not map without mismatches (we used $k$=17). Based on this idea, we generated 10 million 100bp random reads, using a uniform random distribution for each of the four nucleotides (i.e., A, C, G, T have probability 1/4). We also built an index of 17-mers from all genomes in the full NCBI/RefSeq database. Using this index, we counted the number of unknown 17-mers in each random read. Then, we stored one million read that contains at least five unknown 17-mers in dataset "LM", one million read that contain exactly four unknown 17-mers in dataset "MH1", and one million read that contain exactly three unknown 17-mers in dataset "MH2".

## Supplementary Note 2: Generating datasets of "unambiguously mapping reads"

In this note, we describe how we identified and removed ambiguously mapped read from the set of reads generated by ART.

**Definitions and notations**
Given a string $x$, let $|x|$ denote its length.

Definitions: In the following definitions, we assume that $k$ is a positive integer (length of the k-mers), $r$ is a read, and $G$ is a genome.

- Given a set of genomes $\{G_1, G_2, ..., G_m\}$, a $k$-mer $T$ is *specific* to $G_i$ if $T$ occurs in $G_i$ (exactly) but $T$ does not occur (exactly) in any other genome $G_j$, when $j \neq i$ (see [10]).
- Given a set $K$ of $k$-mers specific to $G$, the number of nucleotides of read $r$ covered by at least one $k$-mer in $K$ is called the *coverage* of $r$ to $G$ which we denote by $cov(r,G)$.
- Given a position $l \in [1,|G|-|r|+1]$, we denote by $M(r,G,l)$ the number of mismatches (Hamming distance) between read $r$ and a substring of $G$ of length $|r|$ starting at position $l$.
- We denote by $OPT(r,G) = \min_{l \in [1,|G|-|r|+1]} \{M(r,G,l)\}$, i.e., the minimum number of mismatches for all possible positions of $r$ in $G$.
- Given a set of genomes $\{G_1, G_2, ..., G_m\}$, read $r$ is *unambiguously mapped* to $G_i$ if and only if for all $j \neq i$ we have that $OPT(r,G_i) < OPT(r,G_j)$. In other words, there is no pair of genomes $(G_i, G_j)$ such that the two optimal alignments of $r$ to $G_i$ and $G_j$ achieves the same number of mismatches.

Lemma: Given a read $r$ and a set of genomes $\{G_1, G_2, ..., G_m\}$, if there exists an index $i \in [1, m]$ and a position $l \in [1,|G_i|-|r|+1]$ such that if $\lfloor cov(r,G_i)/k \rfloor > M(r,G_i,l)$ then for all $j \neq i$, we have that $OPT(r,G_j) > OPT(r,G_i)$.
*Proof*: By the definition of $k$-mer specific to a genome: for each non-overlapping block $B$ of $k$ nucleotides that are covered by at least one $k$-mer specific to $G_i$ in $r$, there exists at least one mismatch between block $B$ and any block of $k$ nucleotides in $G_j$ where $i \neq j$. Since there is at least $\lfloor cov(r,G_i)/k \rfloor$ non-overlapping block(s) of $k$ nucleotides covered by at least one $k$-mer $G_i$-specific in $r$, for all $j \neq i$ we have that $OPT(r,G_j) \geq \lfloor cov(r,G_i)/k \rfloor$. By the definition of $OPT$, we have that $OPT(r,G_i) \leq M(r,G_i,l)$. Then, for all $j \neq i$, $OPT(r,G_j) \geq \lfloor cov(r,G_i)/k \rfloor$ and, by the hypothesis of the lemma, we have that $\lfloor cov(r,G_i)/k \rfloor > M(r,G_i,l)$ which implies that $OPT(r,G_j) \geq \lfloor cov(r,G_i)/k \rfloor > M(r,G_i,l) \geq OPT(r,G_i)$. Thus, for all $j \neq i$, we have that $OPT(r,G_j) > OPT(r,G_i)$.

In other words, if $\lfloor cov(r,G_i)/k \rfloor$ is higher than the number of mismatches between $r$ and $G_i$ then read $r$ is unambiguously mapped to $G_i$.

Generating unambiguously mapped reads: We used the ART read simulator to create simulated datasets. We considered the species rank, so genomes of the same species were considered together as a unique sequence. We set $k$=19 to determine sets of $k$-mers specific to each species (i.e., 14,675 sets), then we created a hash-table to extract all 19-mers from all species and remove all 19-mers that are common to at least one pair of species. To create a dataset of unambiguously mapped reads, we filtered reads as follows. For each species $G$ of a given dataset, and for each read $r$ created, we use the alignment (provided by ART) of $r$ to its reference sequence of origin. We compute the number of mismatches $M$ between $r$ and $G$, and we estimated the specificity-coverage $C$ of $r$ to $G$. Using the previous Lemma, $r$ was added to the filtered variant of the dataset (because it is unambiguously mapped to $G$) if the value $C/k$ was higher than $M$+1.

In this step, we are trying to address the issue of reads that were generated from a genome $A$ but could also occur in another genome $B$. If a tool assigns those reads to $B$, should this be considered an incorrect classification or not? The amount of ambiguity depends not only on the dataset, but also on the set of reference genomes used to classify. This ambiguity introduces a hidden dataset- and reference-dependent variable that affects precision and sensitivity. While we understand that these additional datasets are not realistic and artificial, removing ambiguous reads allow us to have an unambiguous "ground truth" that allows to compare across tools without a possible bias.

**Supplementary Note 3: Multithreading algorithm**

To process large FASTQ/FASTA files in parallel, in a memory-scalable fashion, CLARK-*S* exploits a multithreading algorithm similar to that of CLARK [10]. For single-end reads, CLARK-*S* extracts a continuous block of reads (up to two million), partitions the block into *n* bins of reads of equal size (where *n* is the number of parallel threads requested by the user), then classifies the reads of each of these bins in parallel and once all threads are completed CLARK-*S* writes the results in disk and repeat this process for the next block of reads, until all reads have been processed. In the case of paired-end reads, the algorithm is identical than for the single-end reads case, except that the two FASTQ files are first merged (i.e., each read pair is concatenated with the spacer "NNNN" in between them). This memory-scalable multithreading algorithm assures that the RAM-usage remains constant independently of the size of the sample file (see Table below). This technique was also applied in the mapping tool for bisulfite-treated reads Brat-nova [5]. The following table describes the performance (RAM usage and speed) of CLARK-*S'* multithreading approach for several values of *n* on the simBA-525 dataset.

| **Number of threads** | 1 | 2 | 4 | 8 |
|---|---|---|---|---|
| **Speed ($10^3$ reads per min)** | 203.1 | 424.8 | 708.4 | 1,092.5 |
| **RAM usage (GB)** | 108 | 108 | 108 | 108 |

**Supplementary Table 1:** Number of reads and species in each synthetic datasets (default and **unambiguous**) and for the negative controls.

| Synthetic datasets | Buc12 | CParMed48 | Gut20 | Hou31 | Hou21 | Soi50 | simBA-525 |
|---|---|---|---|---|---|---|---|
| **Species** | 12 | 48 | 20 | 31 | 21 | 50 | 525 |
| **Reads (default)** | 600,000 | 1,200,000 | 500,000 | 775,000 | 525,000 | 2,500,000 | 5,666,143 |
| **Reads (unambiguous)** | 600,000 | 1,200,000 | 500,000 | 750,000 | 500,000 | 2,500,000 | 5,727,654 |

| Negative control | HM1 | HM2 | LM |
|---|---|---|---|
| **Reads** | 1,000,000 | 1,000, 000 | 1,000,000 |

**Supplementary Table 2:** Precision and sensitivity for Kraken, CLARK, and CLARK-*S* on the synthetic datasets (default, **unambiguous**). The highest value for precision and sensitivity are indicated in bold. The second table reports the count of classified reads for Kraken, CLARK and CLARK-*S* for the negative controls.

| **Synthetic datasets** | **Kraken** | | **CLARK** | | **CLARK-*S*** | |
|---|---|---|---|---|---|---|
| **Default** | ***Precision*** | ***Sensitivity*** | ***Precision*** | ***Sensitivity*** | ***Precision*** | ***Sensitivity*** |
| **Buc12** | 93.43% | 69.42% | **93.61%** | 69.05% | 90.36% | **71.38%** |
| **CParMed48** | 99.08% | 92.31% | **99.09%** | 92.18% | 99.08% | **93.15%** |
| **Gut20** | 99.21% | 82.45% | **99.24%** | 82.23% | 98.19% | **86.06%** |
| **Hou31** | 94.25% | 83.46% | **94.30%** | 83.30% | 93.94% | **84.32%** |
| **Hou21** | 98.66% | 87.00% | **98.72%** | 86.81% | 98.51% | **88.30%** |
| **Soi50** | 99.49% | 92.48% | **99.51%** | 92.37% | 99.32% | **93.51%** |
| **simBA-525** | 91.17% | 57.57% | **91.27%** | 57.19% | 87.50% | **58.53%** |
| **Unambiguous** | ***Precision*** | ***Sensitivity*** | ***Precision*** | ***Sensitivity*** | ***Precision*** | ***Sensitivity*** |
| **Buc12** | 95.02% | 73.18% | **95.26%** | 72.82% | 92.67% | **75.61%** |
| **CParMed48** | 99.50% | 94.07% | 99.51% | 93.91% | **99.64%** | **95.18%** |
| **Gut20** | 98.87% | 84.82% | **98.92%** | 84.60% | 98.68% | **86.06%** |
| **Hou31** | 97.26% | 87.57% | **97.36%** | 87.45% | 97.09% | **88.21%** |
| **Hou21** | 99.16% | 87.12% | 99.19% | 86.88% | **99.27%** | **89.23%** |
| **Soi50** | 99.49% | 92.96% | **99.51%** | 92.86% | 99.44% | **93.66%** |
| **simBA-525** | 98.57% | 88.75% | **98.69%** | 88.63% | 98.43% | **89.20%** |

| **Negative control** | **Kraken** | **CLARK** | **CLARK-*S*** |
|---|---|---|---|
| **MH1** | 0 | 0 | 0 |
| **MH2** | 0 | 0 | 0 |
| **LM** | 0 | 0 | 0 |

**Supplementary Table 3:** Metadata of the selected real samples from [2]: Sample ID, number of raw reads, number of reads after trimming, object swabbed, location of the sample, borough name, and the number of weekly riders in 2013. Raw reads were trimmed as done in [2]: the first/last 10bp each read were removed (reads longer than 100bp were truncated and the first 100bp were kept); trimmed reads with more than 10 bases with quality scores less than 20 were removed.

| Sample ID | *Raw reads* | *Trimmed reads* | *Object swabbed* | *Location* | *Borough* | *Weekly riders* |
|---|---|---|---|---|---|---|
| **GC01** | 29,282,945 | 28,739,916 | Water Sample | Gowanus Canal | Brooklyn | NA |
| **P00090** | 3,161,196 | 3,085,871 | Stairway rail | Times Sq-42 St/42 St | Manhattan | 197,696 |
| **P00302** | 12,206,080 | 11,700,388 | Bench | 59 St-Columbus Circle | Manhattan | 72,236 |
| **P00306** | 7,536,640 | 7,194,993 | Kiosk | 34 St-Penn Station | Manhattan | 90,042 |
| **P00454** | 7,872,512 | 7,555,783 | Bench | Fulton St | Manhattan | 64,461 |
| **P00589** | 3,129,344 | 3,015,949 | Turnstile | Broadway-Lafayette St/Bleecker St | Manhattan | 38,799 |
| **P00720** | 6,833,000 | 6,536,830 | Bench | Franklin St | Manhattan | 5,825 |
| **P00945** | 7,530,914 | 7,257,415 | Bench | Forest Av | Queens | 4,103 |
| **P01041** | 1,171,456 | 1,160,282 | Bench | Van Siclen Av | Brooklyn | 2,974 |
| **P01136** | 6,417,114 | 6,220,889 | Garbage Can | Jefferson St | Brooklyn | 6,612 |
| **P01270** | 17,072,185 | 16,471,331 | Seats | F Train | Brooklyn | NA |
| **P01324** | 2,686,976 | 2,594,672 | Garbage Can | Whitlock Av | Bronx | 1,685 |

**Supplementary Table 4:** List of species detected in [2] which are also present in the database (i.e., bacteria/archaea/viruses genomes from NCBI/RefSeq) for each of the twelve samples.

| Sample ID | Species in [2] and present in the default RefSeq database (bacteria/archaea/viruses) |
|---|---|
| **GC01** | *Bifidobacterium adolescentis, Bifidobacterium longum, Desulfobacterium autotrophicum, Erwinia billingiae, Eubacterium eligens, Eubacterium rectale, Methanocorpusculum labreanum, Parabacteroides distasonis* |
| **P00090** | *Acinetobacter baumannii, Cronobacter turicensis, Enterobacter cloacae, Enterococcus casseliflavus, Enterococcus faecalis, Klebsiella pneumoniae, Lysinibacillus sphaericus, Macrococcus caseolyticus, Micrococcus luteus, Pseudomonas putida, Pseudomonas stutzeri, Stenotrophomonas maltophilia, Streptococcus suis* |
| **P00302** | *Achromobacter xylosoxidans, Acinetobacter baumannii, Bacillus megaterium, Dickeya dadantii, Enterobacter cloacae, Enterococcus casseliflavus, Enterococcus faecalis, Enterococcus faecium, Enterococcus hirae, Finegoldia magna, Klebsiella pneumoniae, Lactococcus lactis, Leuconostoc mesenteroides, Lysinibacillus sphaericus, Micrococcus luteus, Propionibacterium acidipropionici, Propionibacterium acnes, Pseudomonas putida, Pseudomonas stutzeri, Staphylococcus epidermidis, Staphylococcus haemolyticus, Stenotrophomonas maltophili* |
| **P00306** | *Acinetobacter baumannii, Acinetobacter oleivorans, Enterobacter cloacae, Enterobacteria phage IME10, Enterococcus casseliflavus, Enterococcus faecium, Klebsiella pneumoniae, Propionibacterium acnes, Pseudomonas stutzeri, Stenotrophomonas maltophilia* |
| **P00454** | *Acinetobacter baumannii, Chlorobium phaeobacteroides, Enterobacter cloacae, Enterococcus casseliflavus, Enterococcus mundtii, Klebsiella pneumoniae, Lysinibacillus sphaericus, Pseudomonas stutzeri, Solibacillus silvestris, Stenotrophomonas maltophilia* |
| **P00589** | *Acinetobacter baumannii, Enterobacter cloacae, Enterobacteria phage HK97, Enterococcus casseliflavus, Lactococcus lactis, Pseudomonas putida, Pseudomonas stutzeri, Streptococcus suis* |
| **P00720** | *Corynebacterium variabile, Enterobacter cloacae, Enterobacteria phage HK97, Enterococcus casseliflavus, Lactococcus lactis, Leuconostoc citreum, Lysinibacillus sphaericus, Pseudomonas stutzeri, Stenotrophomonas maltophilia* |
| **P00945** | *Bacillus megaterium, Enterobacter cloacae, Enterococcus faecalis, Enterococcus faecium, Lysinibacillus sphaericus, Pseudomonas putida, Pseudomonas stutzeri, Stenotrophomonas maltophilia, Stenotrophomonas phage phiSMA7* |
| **P01041** | *Enterobacter cloacae, Enterobacteria phage HK97, Enterococcus casseliflavus, Enterococcus faecalis, Pseudomonas stutzeri, Stenotrophomonas maltophilia* |
| **P01136** | *Brucella ovis, Corynebacterium variabile, Enterobacter cloacae, Enterobacteria phage HK97, Enterococcus casseliflavus, Leuconostoc mesenteroides, Pseudomonas putida, Pseudomonas stutzeri, Stenotrophomonas maltophilia, Streptococcus suis* |
| **P01270** | *Achromobacter xylosoxidans, Enterobacter cloacae, Enterococcus casseliflavus, Enterococcus faecalis, Enterococcus faecium, Enterococcus hirae, Lactococcus lactis, Lysinibacillus sphaericus, Propionibacterium acnes, Pseudomonas putida, Pseudomonas stutzeri, Stenotrophomonas maltophilia* |
| **P01324** | *Cronobacter sakazakii, Enterobacter cloacae, Enterobacteria phage HK97, Enterococcus casseliflavus, Enterococcus faecium, Escherichia coli, Klebsiella pneumoniae, Kocuria rhizophila, Lactococcus lactis, Leuconostoc mesenteroides, Micrococcus luteus, Pseudomonas stutzeri, Rhodopseudomonas palustris, Stenotrophomonas maltophilia, Stenotrophomonas phage phiSMA7, Streptococcus parauberis, Streptococcus suis, Streptococcus thermophilus* |

**Supplementary Table 5:** Column A lists the reads count reported by Kraken, CLARK, and CLARK-*S* on the species listed in Supplementary Table 4. For each species, a count is reported as a triplet *(Kraken, CLARK, CLARK-S)*. Column B reports the agreement rate between [2] and results reported by Kraken, CLARK, and CLARK-*S*, in this order. For example, for the sample GC01, the agreement rate between Kraken and [2] was 75% because Kraken detected the presence of 6 species out of the 8 in [2]. Values in bold indicate the highest agreement rate. Column C reports the percentage of species for which CLARK-*S* reports a higher reads count than both Kraken and CLARK. For example, for the sample P00090, CLARK-*S* reports a higher number of reads count than both Kraken and CLARK for 12 species out of 13 (i.e., 92.3%).

| Sample ID | A | B | C |
|---|---|---|---|
| **GC01** | *Bifidobacterium adolescentis (1238, 1218, 1307), Bifidobacterium longum (1106, 1093, 1217), Desulfobacterium autotrophicum (88171, 84690, 142189), Erwinia billingiae (8774, 8651, 9443), Eubacterium eligens (0, 0, 0), Eubacterium rectale (0, 0, 0), Methanocorpusculum labreanum (429, 400, 1091), Parabacteroides distasonis (1028, 1011, 1340)* | **75%, 75%, 75%** | 100% |
| **P00090** | *Acinetobacter baumannii (8482, 8143, 14783), Cronobacter turicensis (2108, 2078, 1471), Enterobacter cloacae (44220, 41877, 64974), Enterococcus casseliflavus (14731, 14535, 16365), Enterococcus faecalis (2481, 2472, 2563), Klebsiella pneumoniae (49647, 49011, 49772), Lysinibacillus sphaericus (4, 4, 11), Macrococcus caseolyticus (1904, 1891, 2110), Micrococcus luteus (2686, 2646, 2990), Pseudomonas putida (8944, 8405, 12327), Pseudomonas stutzeri (1243301, 1228384, 1349618), Stenotrophomonas maltophilia (15162, 14732, 19712), Streptococcus suis (26495, 25484, 41016)* | **100%, 100%, 100%** | 92.3% |
| **P00302** | *Achromobacter xylosoxidans (417007, 396787, 798804), Acinetobacter baumannii (53782, 51650, 84481), Bacillus megaterium (1291, 1263, 1619), Dickeya dadantii (8574, 8893, 6470), Enterobacter cloacae (328816, 303503, 497288), Enterococcus casseliflavus (9735, 9517, 12275), Enterococcus faecalis (20903, 20844, 21109), Enterococcus faecium (773, 757, 1045), Enterococcus hirae (1506, 1500, 1557), Finegoldia magna (314, 305, 505), Klebsiella pneumoniae (32826, 30878, 31901), Lactococcus lactis (911, 873, 1483), Leuconostoc mesenteroides (1890, 1853, 1965), Lysinibacillus sphaericus (1, 1, 1), Micrococcus luteus (781, 785, 879), Propionibacterium acidipropionici (379, 385, 413), Propionibacterium acnes (770, 767, 812), Pseudomonas putida (3493, 3452, 4770), Pseudomonas stutzeri (987112, 980445, 1011820), Staphylococcus epidermidis (661, 650, 771), Staphylococcus haemolyticus (1066, 1028, 1320), Stenotrophomonas maltophilia (50279, 48597, 72008)* | **100%, 100%, 100%** | 86.4% |
| **P00306** | *Acinetobacter baumannii (540511, 520987, 731225), Acinetobacter oleivorans (67230, 66304, 72904), Enterobacter cloacae (171685, 159913, 272355), Enterobacteria phage IME10 (0, 0, 0), Enterococcus casseliflavus (54313, 53029, 67794), Enterococcus faecium (2675, 2649, 2910), Klebsiella pneumoniae (20732, 19474, 22448), Propionibacterium acnes (931, 925, 948), Pseudomonas stutzeri (533478, 525799, 585020), Stenotrophomonas maltophilia (564888, 560201, 586129)* | **90%, 90%, 90%** | 100% |
| **P00454** | *Acinetobacter baumannii (46223, 45761, 48612), Chlorobium phaeobacteroides (1, 1, 147), Enterobacter cloacae (21652, 20137, 32217), Enterococcus casseliflavus (6931, 6852, 7405), Enterococcus mundtii (1112, 1101, 1151), Klebsiella pneumoniae (22895, 22507, 22950), Lysinibacillus sphaericus (1, 1, 3), Pseudomonas stutzeri (4711283, 4652107, 5004594), Solibacillus silvestris (2555, 2407, 4990), Stenotrophomonas maltophilia (43004, 41930, 53308)* | **100%, 100%, 100%** | 100% |

| | | | |
|---|---|---|---|
| **P00589** | *Acinetobacter baumannii (7513, 7362, 9684), Enterobacter cloacae (2471, 2380, 3334), Enterobacteria phage HK97 (0, 0, 10), Enterococcus casseliflavus (11906, 11742, 13533), Lactococcus lactis (1743, 1699, 2578), Pseudomonas putida (6062, 5822, 8554), Pseudomonas stutzeri (777233, 765277, 850289), Streptococcus suis (8506, 8201, 13373)* | 87.5%, 87.5%, **100%** | 100% |
| **P00720** | *Corynebacterium variabile (1302, 1262, 1487), Enterobacter cloacae (82530, 75880, 125426), Enterobacteria phage HK97 (0, 0, 48), Enterococcus casseliflavus (25280, 25059, 26621), Lactococcus lactis (2437, 2430, 2614), Leuconostoc citreum (498, 496, 511), Lysinibacillus sphaericus (26, 25, 49), Pseudomonas stutzeri (2738041, 2698911, 2989300), Stenotrophomonas maltophilia (516748, 501500, 671902)* | 88.9%, 88.9%, **100%** | 100% |
| **P00945** | *Bacillus megaterium (760, 754, 771), Enterobacter cloacae (44780, 41433, 69336), Enterococcus faecalis (8984, 8954, 9128), Enterococcus faecium (1219, 1217, 1278), Lysinibacillus sphaericus (2, 0, 2), Pseudomonas putida (2505, 2340, 2920), Pseudomonas stutzeri (4149, 4157, 4849), Stenotrophomonas maltophilia (1258848, 1230418, 1589727), Stenotrophomonas phage phiSMA7 (397, 391, 637)* | **100%**, 88.9%, **100%** | 100% |
| **P01041** | *Enterobacter cloacae (13726, 12754, 20206), Enterobacteria phage HK97 (0, 0, 11), Enterococcus casseliflavus (5196, 5082, 6395), Enterococcus faecalis (2571, 2567, 2607), Pseudomonas stutzeri (611583, 608607, 626318), Stenotrophomonas maltophilia (58910, 58591, 60892)* | 83.3%, 83.3%, **100%** | 100% |
| **P01136** | *Brucella ovis (0, 0, 12), Corynebacterium variabile (974, 965, 1005), Enterobacter cloacae (41486, 38925, 60976), Enterobacteria phage HK97 (0, 0, 16), Enterococcus casseliflavus (8871, 8783, 9460), Leuconostoc mesenteroides (896, 886, 909), Pseudomonas putida (49887, 47305, 56607), Pseudomonas stutzeri (1140608, 1101902, 1627874), Stenotrophomonas maltophilia (6588, 6425, 9192), Streptococcus suis (7045, 6768, 10659)* | 80%, 80%, **100%** | 100% |
| **P01270** | *Achromobacter xylosoxidans (9129, 9013, 10142), Enterobacter cloacae (464185, 438737, 712806), Enterococcus casseliflavus (204915, 203223, 215280), Enterococcus faecalis (454647, 453560, 458843), Enterococcus faecium (5058, 4972, 6434), Enterococcus hirae (7299, 7264, 7588), Lactococcus lactis (2155, 2119, 2684), Lysinibacillus sphaericus (7, 6, 12), Propionibacterium acnes (341, 366, 351), Pseudomonas putida (1722194, 1623230, 3097829), Pseudomonas stutzeri (3177433, 3126518, 3511417), Stenotrophomonas maltophilia (1281605, 1248952, 1619141)* | **100%, 100%, 100%** | 91.7% |
| **P01324** | *Cronobacter sakazakii (4237, 4016, 4891), Enterobacter cloacae (15067, 13986, 22082), Enterobacteria phage HK97 (0, 0, 2), Enterococcus casseliflavus (4685, 4553, 6638), Enterococcus faecium (533, 514, 783), Escherichia coli (2797, 2694, 4119), Klebsiella pneumoniae (2859, 2702, 3091), Kocuria rhizophila (84, 70, 178), Lactococcus lactis (1088, 1071, 1322), Leuconostoc mesenteroides (1042, 1036, 1089), Micrococcus luteus (162, 166, 173), Pseudomonas stutzeri (323280, 319408, 343408), Rhodopseudomonas palustris (370, 354, 422), Stenotrophomonas maltophilia (72640, 70301, 105826), Stenotrophomonas phage phiSMA7 (2, 2, 4), Streptococcus parauberis (1477, 1473, 1526), Streptococcus suis (378, 359, 582), Streptococcus thermophiles (369, 367, 389)* | 94.4%, 94.4%, **100%** | 100% |

**Supplementary Table 6:** Assignment rate (i.e., ratio in percent between the number of assigned/classified reads and the total number of reads) on real samples for Kraken, CLARK and CLARK-*S*. Values in bold are the highest.

| Sample ID | Kraken | CLARK | CLARK-*S* |
|---|---|---|---|
| **GC01** | 1.74% | 1.36% | **2.55%** |
| **P00090** | 54.22% | 49.59% | **56.16%** |
| **P00302** | 29.07% | 23.70% | **29.89%** |
| **P00306** | 39.37% | 33.82% | **40.47%** |
| **P00454** | 70.02% | 66.37% | **71.50%** |
| **P00589** | 31.84% | 29.46% | **34.24%** |
| **P00720** | 59.49% | 55.59% | **64.35%** |
| **P00945** | 26.26% | 23.21% | **35.65%** |
| **P01041** | **67.87%** | 50.28% | 64.35% |
| **P01136** | 31.01% | 26.36% | **35.65%** |
| **P01270** | **65.20%** | 50.28% | 64.35% |
| **P01324** | **27.65%** | 23.29% | 27.23% |

**Supplementary Table 7:** Classification speed of Kraken, CLARK and CLARK-*S* on the synthetic datasets (default and **unambiguous**), the negative control samples and the real samples. CLARK and Kraken were run with default settings (i.e., 31-mers), and, for Kraken, the database was loaded with the option "—preload" to assure the highest speed. Each tool was run three times to smooth I/O and cache issues (the reported numbers are the best values). The values are in thousands of read per minute. Values in bold are the highest for each dataset.

| **Default** | **Kraken (1 CPU)** | **CLARK (1 CPU)** | **CLARK-*S* (1 CPU)** | **CLARK-*S* (8 CPUs)** |
|---|---|---|---|---|
| **Buc12** | 2,206.0 | **4, 839.5** | 214.4 | 1, 220.8 |
| **CParMed48** | 2,060.9 | **3, 691.4** | 204.3 | 913.6 |
| **Gut20** | 1,792.6 | **3, 369.5** | 196.1 | 1, 077.8 |
| **Hou31** | 2,111.6 | **3, 465.5** | 201.4 | 1, 067.7 |
| **Hou21** | 2,011.5 | **3, 308.9** | 199.2 | 1, 124.6 |
| **Soi50** | 2,008.6 | **3, 193.3** | 169.5 | 1, 074.7 |
| **simBA-525** | 1,955.7 | **3, 194.5** | 203.1 | 1, 092.5 |
| **Unambiguous** | **Kraken (1 CPU)** | **CLARK (1 CPU)** | **CLARK-*S* (1 CPU)** | **CLARK-*S* (8 CPUs)** |
| **Buc12** | 2,307.8 | **4, 160.5** | 217.7 | 1, 101.5 |
| **CParMed48** | 2,299.3 | **4, 057.7** | 201.3 | 874.1 |
| **Gut20** | 2,028.0 | **2, 954.0** | 134.3 | 1, 083.7 |
| **Hou31** | 2,109.3 | **3, 912.9** | 142.0 | 964.0 |
| **Hou21** | 2,057.8 | **3, 801.1** | 157.8 | 1, 003.8 |
| **Soi50** | 2,131.6 | **2, 868.9** | 141.4 | 1, 024.7 |
| **simBA-525** | 1,936.1 | **3, 359.0** | 141.7 | 1, 076.3 |

| **Negative control** | **Kraken (1 CPU)** | **CLARK (1 CPU)** | **CLARK-*S* (1 CPU)** | **CLARK-*S* (8 CPUs)** |
|---|---|---|---|---|
| **HM1** | 1,924.7 | **2, 619.1** | 146.2 | 1, 033.1 |
| **HM2** | 1,901.6 | **2, 932.1** | 131.9 | 937.9 |
| **LM** | 2,145.8 | **2, 654.2** | 134.2 | 957.3 |

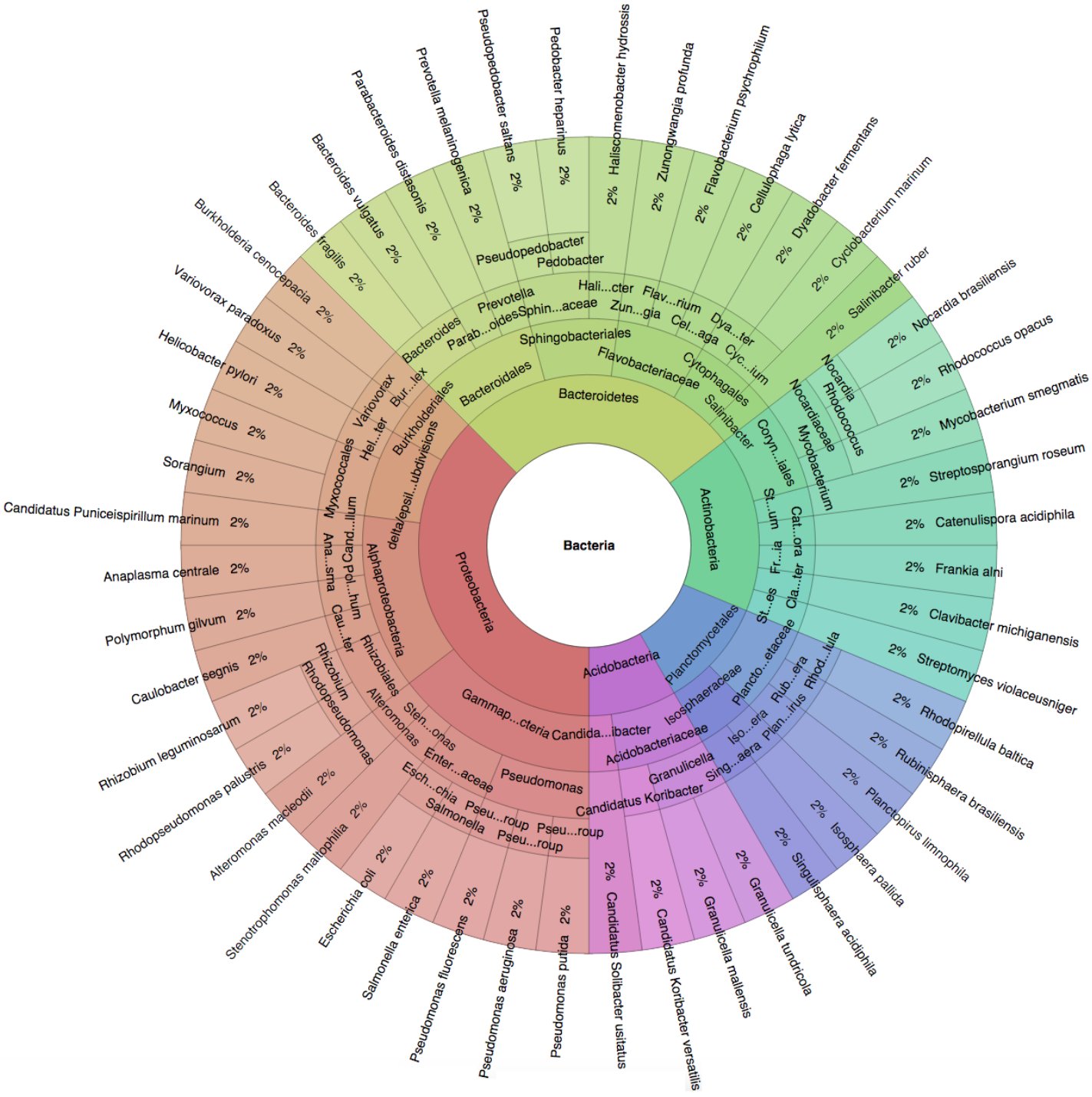
| **Sample ID** | **Kraken (1 CPU)** | **CLARK (1 CPU)** | **CLARK-*S* (1 CPU)** | **CLARK-*S* (8 CPUs)** |
|---|---|---|---|---|
| **GC01** | 2,572.8 | **3,142.3** | 290.7 | 1,315.9 |
| **P00090** | 2,543.3 | **2,587.7** | 230.7 | 1,355.7 |
| **P00302** | 2,310.9 | **3,330.3** | 326.7 | 1,432.1 |
| **P00306** | 2,596.5 | **3,553.6** | 332.5 | 1,428.1 |
| **P00454** | 2,709.9 | **3,668.7** | 364.7 | 1,569.5 |
| **P00589** | 2,805.0 | **4,929.9** | 312.2 | 1,373.8 |
| **P00720** | 2,457.0 | **5,203.0** | 312.2 | 1,545.8 |
| **P00945** | 2,683.1 | **4,758.7** | 324.2 | 1,390.9 |
| **P01041** | 2,311.6 | **4,348.5** | 313.9 | 1,381.2 |
| **P01136** | 2,643.1 | **4,893.1** | 315.0 | 1,371.2 |
| **P01270** | 2,390.5 | **3,548.8** | 341.8 | 1,531.8 |
| **P01324** | 2,660.8 | **3,513.6** | 320.1 | 1,363.9 |

**Supplementary Table 8:** Memory usage and running time for the index creation and classification for Kraken, CLARK and CLARK-S. The database is the bacterial, archaeal and viral sequences from NCBI/RefSeq. Measures indicated were obtained via the "`/usr/bin/time -v`" command. All tools (Kraken v0.10-5, CLARK/CLARK-S v1.2.3) were run on a Linux server (20 cores Intel Xeon CPU E5-2690v2 3.3GHz and 512GB of RAM). Lowest values are indicated in bold.

| | | **Kraken** | **CLARK** | **CLARK-*S*** |
|---|---|---|---|---|
| **Index creation** | Memory usage (1 CPU) | 160 Gb | **156 Gb** | **156 Gb** |
| | Total running time (1 CPU) | 6h50m | **3h20m** | 9h40m |
| | Database space in disk | 152 Gb | **34 Gb** | 101 Gb |
| **Classification** | Memory usage (1 CPU) | 79 Gb | **58 Gb** | 108 Gb |

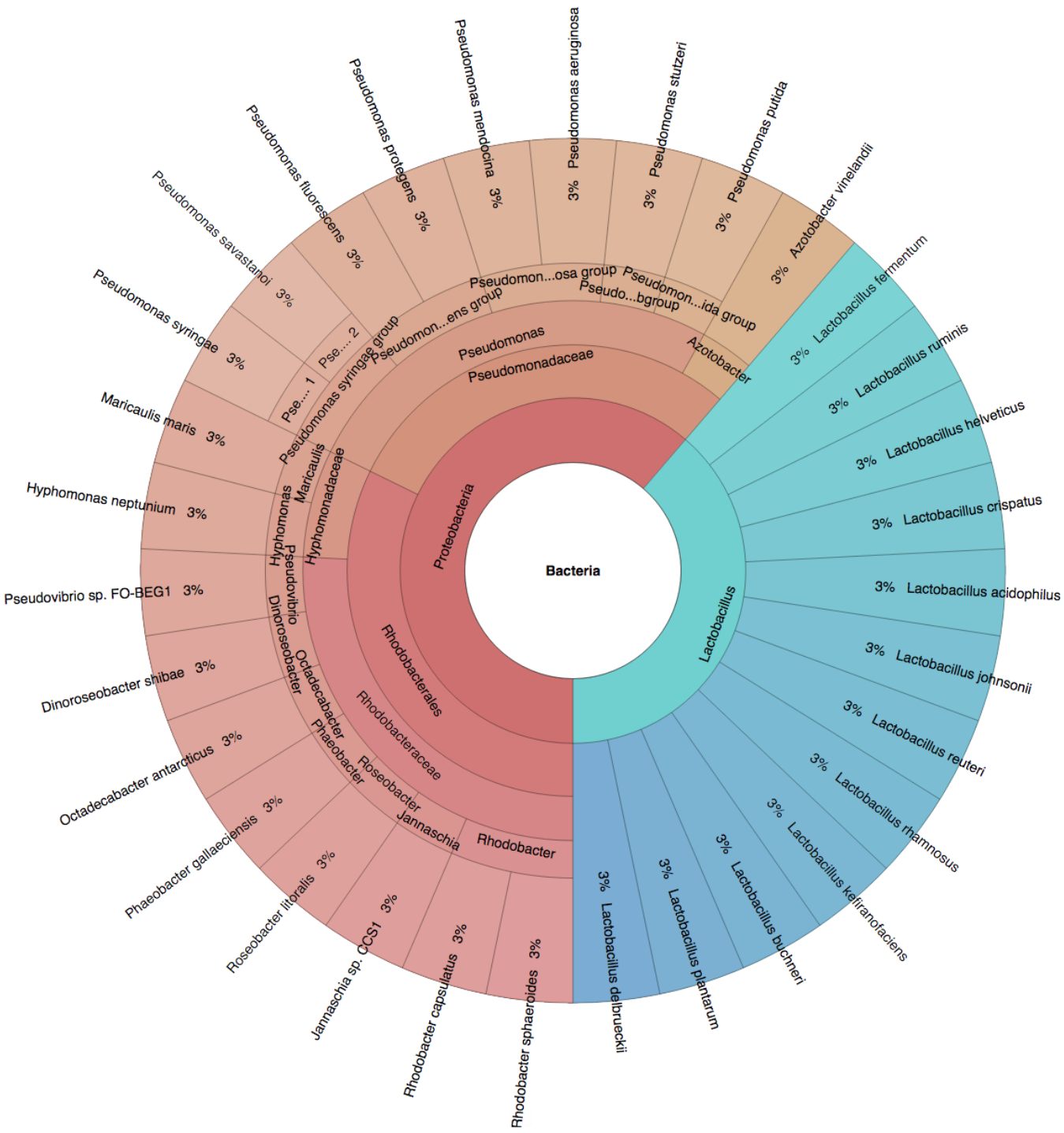
**Supplementary Figure 1: Buc12**

**Supplementary Figure 2: CParMed48**



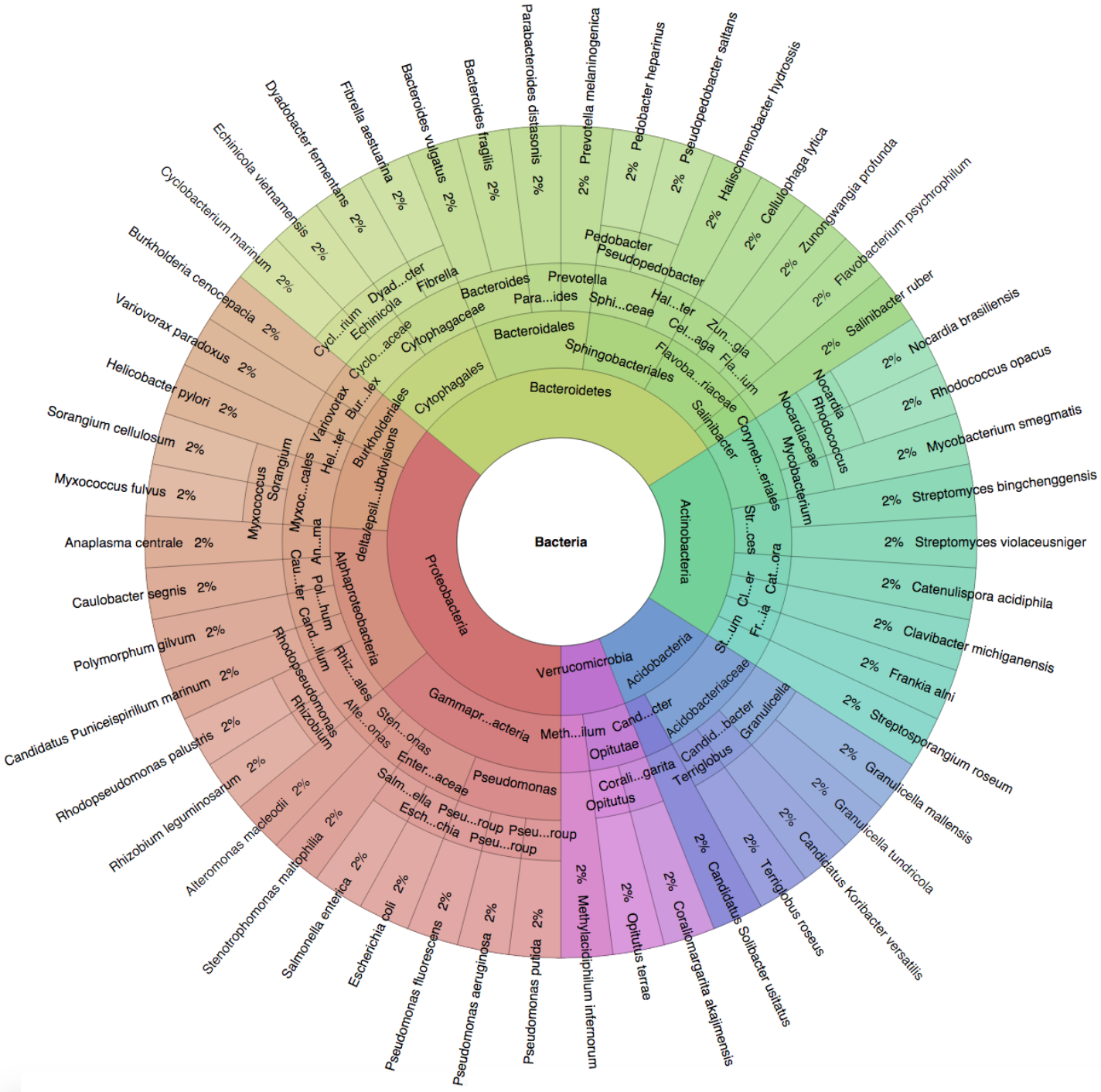
**Supplementary Figure 3: Gut20**

**Supplementary Figure 4: Hous31**

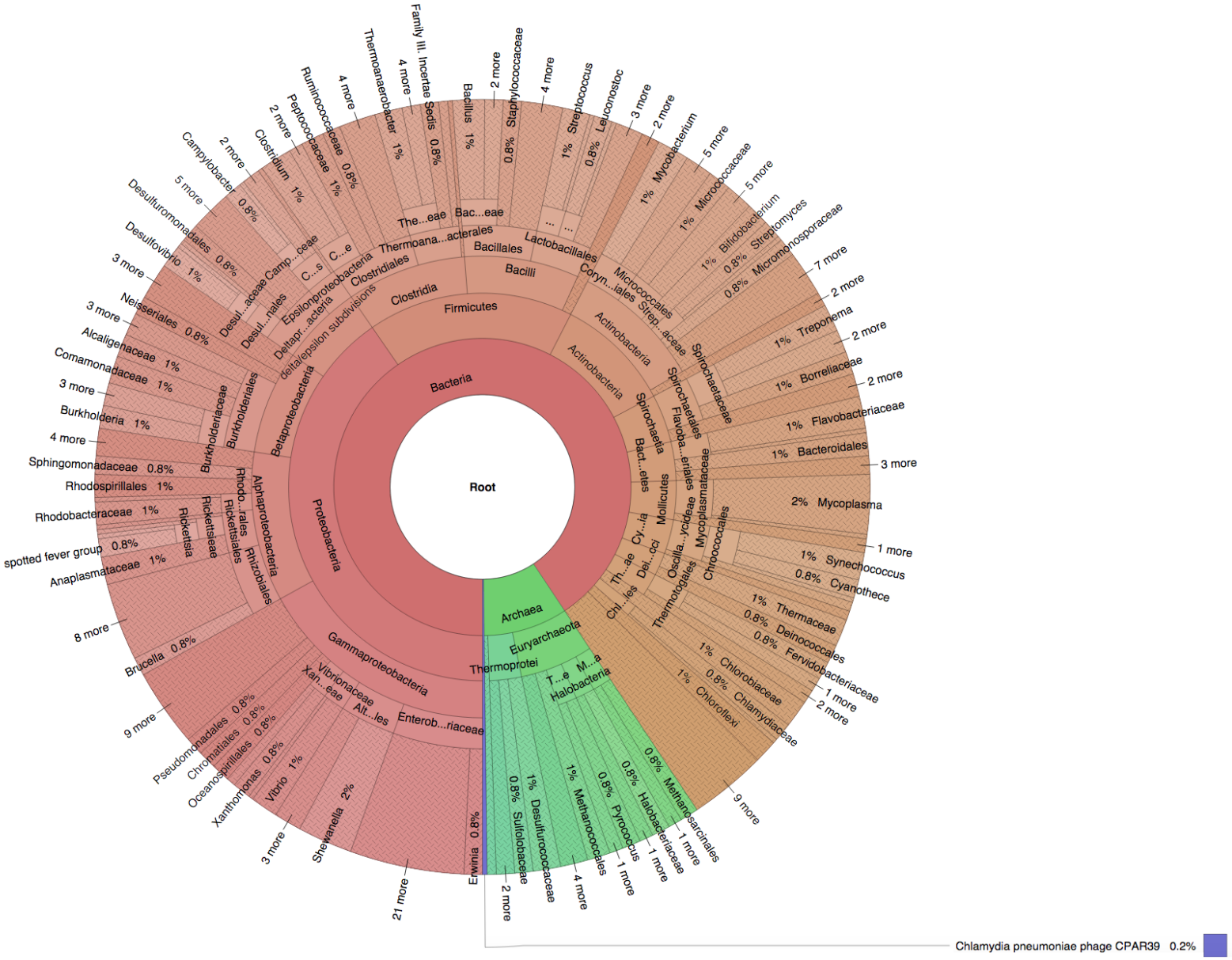


**Supplementary Figure 5: Hous21**

**Supplementary Figure 6: Soi50**



**Supplementary Figure 7: simBA-525**



## References


[1] Adams, R. I., Bateman, A. C., Bik, H. M., & Meadow, J. F. (2015). Microbiota of the indoor environment: a meta-analysis. *Microbiome*, *3*(1), 1.

[2] Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J. M., Reeves, D., Gandara, J., Chhangawala, S., et al. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell systems*, *1*(1), 72-87.

[3] Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., and Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390-21395.

[4] Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J., & Huttenhower, C. (2015). Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, *112*(22), E2930-E2938.

[5] Harris, E. Y., Ounit, R., and Lonardi, S. (2016). BRAT-nova: Fast and accurate mapping of bi-sulfite-treated reads. *Bioinformatics*, page btw226.

[6] Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207-214.

[7] Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593-594.

[8] Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., & Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature biotechnology*, *34*(1), 64-69.

[9] Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, *12*(1), 1.

[10] Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, *16*(1), 1.

[11] Reese, A. T., Savage, A., Youngsteadt, E., McGuire, K. L., Koling, A., Watkins, O., Frank, S. D., and Dunn, R. R. (2015). Urban stress is associated with variation in microbial species composition—but not richness—in Manhattan. *The ISME journal*.

[12] Ruiz-Calderon, J. F., Cavallin, H., Song, S. J., Novoselac, A., Pericchi, L. R., Hernandez, J. N., Rios, R., Branch, O. H., Pereira, H., Paulino, L. C., et al. (2016). Walls talk: Microbial biogeography of homes spanning urbanization. *Science advances*, *2*(2), e1501061.