Lu and Salzberg *Microbiome* (2020) 8:124
https://doi.org/10.1186/s40168-020-00900-2

Microbiome

RESEARCH Open Access

Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2

Jennifer Lu[1,2*] and Steven L. Salzberg[1,2,3]

Abstract

Background: For decades, 16S ribosomal RNA sequencing has been the primary means for identifying the bacterial species present in a sample with unknown composition. One of the most widely used tools for this purpose today is the QIIME (Quantitative Insights Into Microbial Ecology) package. Recent results have shown that the newest release, QIIME 2, has higher accuracy than QIIME, MAPseq, and mothur when classifying bacterial genera from simulated human gut, ocean, and soil metagenomes, although QIIME 2 also proved to be the most computationally expensive. Kraken, first released in 2014, has been shown to provide exceptionally fast and accurate classification for shotgun metagenomics sequencing projects. Bracken, released in 2016, then provided users with the ability to accurately estimate species or genus relative abundances using Kraken classification results. Kraken 2, which matches the accuracy and speed of Kraken 1, now supports 16S rRNA databases, allowing for direct comparisons to QIIME and similar systems.

Methods: For a comprehensive assessment of each tool, we compare the computational resources and speed of QIIME 2's q2-feature-classifier, Kraken 2, and Bracken in generating the three main 16S rRNA databases: Greengenes, SILVA, and RDP. For an evaluation of accuracy, we evaluated each tool using the same simulated 16S rRNA reads from human gut, ocean, and soil metagenomes that were previously used to compare QIIME, MAPseq, mothur, and QIIME 2. We evaluated accuracy based on the accuracy of the final genera read counts assigned by each tool. Finally, as Kraken 2 is the only tool providing per-read taxonomic assignments, we evaluate the sensitivity and precision of Kraken 2's per-read classifications.

Results: For both the Greengenes and SILVA database, Kraken 2 and Bracken are up to 100 times faster at database generation. For classification, using the same data as previous studies, Kraken 2 and Bracken are up to 300 times faster, use 100x less RAM, and generate results that more accurate at 16S rRNA profiling than QIIME 2's q2-feature-classifier.

Conclusion: Kraken 2 and Bracken provide a very fast, efficient, and accurate solution for 16S rRNA metataxonomic data analysis.

*Correspondence: jennifer.lu717@gmail.com
[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
Full list of author information is available at the end of the article

Introduction

Since the 1970s, sequencing of the 16S ribosomal RNA gene has been used for analyzing and identifying bacterial communities [1, 2]. This technology targets the 16S rRNA gene, which has regions that are both highly conserved and highly variable (hypervariable) among bacterial species. The highly conserved regions allow for the design of "universal" PCR primers to target and amplify the 16S rRNA sequence, while the hypervariable regions allow for discrimination among different bacterial clades. These properties allow 16S rRNA sequencing experiments to capture nearly all of the bacteria in a microbial community, which can then be compared to large 16S rRNA databases to determine their identities.

Researchers have utilized 16S rRNA sequencing for a very broad range of environmental and clinical studies. For example, the Earth Microbiome Project [3] and other environmental studies have used 16S rRNA sequencing to reveal the bacterial diversity of soil [4, 5], beach sand [6], and ocean environments [7], while other studies targeted the microbiome of plants [8–10]. In the clinic, 16S rRNA has been used for diagnostic purposes to identify infectious bacterial species [11–13] and to characterize the role of bacterial diversity in diseases such as diabetes [14], Alzheimer's disease [15], cancer [16], and autism [17]. The Human Microbiome Project, along with other human microbiome studies, has used 16S rRNA data to characterize the bacterial community present in the human gut, feces, skin, and other areas of the body [18–20].

16S rRNA classification

Analysis of the bacterial community from a 16S rRNA sequencing experiment includes comparing the reads to reference database. The tool most widely used for 16S rRNA analysis and classification today is the Quantitative Insights into Microbial Ecology (QIIME) software package [21], which compares sequencing reads against a 16S rRNA reference database. The three standard 16S rRNA databases, each of which has somewhat different content, are Greengenes [22], SILVA [23], and RDP [24].

First released in 2011, QIIME 1 [21] provided 4 classification algorithms for 16S rRNA, respectively based on the RDP classifier [25], BLAST [26], UCLUST [27], and SortMeRNA [28]. In 2018, QIIME 2 [29]'s q2-feature-classifier was released [30], adding 3 new classification algorithms based on scikit-learn's naïve Bayes algorithm [31], VSEARCH [32], and BLAST+ [33]. By default, QIIME 1 uses the UCLUST algorithm for classification while QIIME 2 suggests usage of the naïve Bayes algorithm.

In 2018, Almeida et al. [34] performed benchmark tests comparing QIIME 2 to its predecessor, QIIME 1, and to two additional 16S rRNA classification tools, MAPseq [35] and mothur [36]. Almeida et al. evaluated the performance of each tool by classifying 16S rRNA reads that were simulated from bacteria known to be present in human gut, soil, and ocean microbiomes. That study concluded that QIIME 2's q2-feature-classifier provides the best accuracy on the basis of recall and F-score. However, they also noted that QIIME 2 was the most computationally expensive, requiring substantially more CPU time and more memory than other tools.

Kraken, Kraken 2, and Bracken

The Kraken program uses an alignment-free algorithm that, when first released in 2014, was hundreds of times faster than any previously described program for shotgun metagenomics sequence analysis, with accuracy comparable to BLAST and superior to other tools [37]. Using a single thread, Kraken can classify sequence data at a rate of $>$ 1 million reads per minute.

In 2016, Bracken was released as an extension to Kraken to estimate species abundance from Kraken's output [38]. As originally designed, Kraken attempts to classify each read as specifically as possible, allowing reads to be classified at any taxonomic level depending on how many genomes share the same sequence. For example, a read that has identical matches to two species will be classified at the genus level. Bracken adds the capability of abundance estimation to Kraken, i.e., using Kraken's read counts and prior knowledge of the database sequences, it estimates read counts for all species, genera, or higher-level taxa in a sample. For example, when Bracken is asked to estimate species counts, it will re-distribute all reads that Kraken assigns at the genus level (or higher) down to the species level.

Kraken 2, released in 2018, implemented significant changes to the database structure and classification steps to make databases smaller and classifications faster [39]. Because it uses the same classification algorithm, Kraken 2 has nearly the same precision and sensitivity as Kraken 1. However, Kraken 2 now also provides direct support for 16S rRNA classification with any of the three standard 16S rRNA databases: Greengenes, SILVA, and RDP. This new feature allowed us to compare Kraken 2 to the current state-of-the-art programs for 16S rRNA classification, as described below.

Kraken 2 versus QIIME 2

In 2016, Lindgreen et al. evaluated 14 metagenomics classifiers, including Kraken 1 and QIIME 1 (UCLUST) [40]. That study showed that Kraken achieved the lowest false positive rate, 0%, while QIIME had a false positive rate of 0.28%. Kraken also had higher sensitivity than QIIME, correctly labeling 70% of the reads while QIIME was correct on 60%. Finally, Kraken obtained a Pearson correlation between the known and predicted relative abundances of phyla and genera of 0.99, versus 0.78

for QIIME. However, that study used different databases and different input data (reads produced by metagenomic shotgun sequencing) to evaluate these tools. For Kraken 1, Lindgreen et al. measured its performance on all input sequences from a shotgun metagenomics experiment, using a database containing all complete bacteria and archaeal genomes from RefSeq, while for QIIME 1, they analyzed its performance only on 16S rRNA sequences against the 16S Greengenes database.

Because QIIME has primarily been used for 16S rRNA sequencing projects and Kraken has previously been used primarily for metagenomics shotgun sequencing projects, the tools have not been directly compared. Here, we compare QIIME 2's q2-feature-classifier and Kraken 2 using the 16S rRNA reads generated in the Almeida et al. benchmark study, using both the Greengenes and SILVA 16S rRNA databases. We also show results for Kraken on the RDP database, which is not compatible with QIIME 2. Because we only tested the most recent version of each tool, we will henceforth refer to QIIME 2 as QIIME and Kraken 2 as Kraken.

Results

Prior to classification, Kraken requires users to first build a specialized database consisting of three files: taxo.k2d, opts.k2d, and hash.k2d. The user also can choose the value *k* that determines the length of the sequences that Kraken uses for its index; every sequence (or *k*-mer) of length *k* is associated with the species in which it occurs. *K*-mers that occur in two or more species are associated with the lowest common ancestor of those species. The database files contain the taxonomy and *k*-mer information for the specified database. Following generation of these files, Bracken requires users to generate a *k*-mer distribution file. Kraken and Bracken additionally allow the use of multiple threads to accelerate database construction. We tested building all files for the 16S rRNA Greengenes 13_8, SILVA 132, and RDP 11.5 databases using 1, 4, 8, and 16 threads. Table 1 summarizes the contents of each of these databases.

For QIIME, users can generate the database (called a "classifier") by first converting sequence and taxonomy files into QIIME compatible qza files. QIIME classifier generation is single-threaded. QIIME does provide prebuilt SILVA and Greengenes taxonomy classifiers for q2-feature-classifier at https://docs.qiime2.org/2020.6/data-resources/. However, to evaluate the classifier generation requirements, we built QIIME naïve-bayes classifiers for Greengenes 13_8 and SILVA 132.

Figure 1a compares the combined database building time for Kraken/Bracken against the classifier generation time of QIIME. Kraken was at least 9x faster than QIIME for database creation, e.g., it took 9 min to build the Greengenes database index, while QIIME required 78 min for the same database. For the SILVA database, Kraken required only 34 min while QIIME required more than 58 h to build the same database. Supplemental File 1 lists all command lines used for building the databases.

To compare the accuracy of Kraken, Bracken, and QIIME, we classified 12 samples generated by Almeida et al. These 12 samples, each containing just under 200,000 reads, represent 3 different metagenomes (human, ocean, and soil) and 4 different 16S rRNA primers (V12, V34, V4, and V45). The number of reads in each sample is shown in Table 2. See the "Methods" section for additional information about sample generation and pre-processing steps.

QIIME classifiers require one single file containing all de-multiplexed reads. Therefore, we provided QIIME with one file per metagenome, each containing reads from all 4 primer sets. However, Kraken and Bracken classify samples one at a time, requiring each of the 12 samples to be processed individually.

Kraken and QIIME provide multi-threading options to speed up classification. We therefore tested Kraken and the QIIME Greengenes classifier using 1, 4, 8, and 16 threads. The QIIME SILVA classifier with 8 threads required approximately 1.5 days of run time, and for this reason, we only tested it using 16 and 8 threads and did not evaluate the QIIME 2 SILVA classifier using 1 or 4 threads.

Figure 1b compares the average time in minutes required by QIIME's q2-feature-classifier vs. Kraken/Bracken to classify a single metagenome using the 16S rRNA Greengenes and SILVA databases. Due to the very large difference in run time between tools, this figure compares the multi-threaded options of QIIME against the single-threaded classification time of Kraken/Bracken. Figure 1c reports the classification times of Kraken/Bracken in seconds.

Another important consideration for software selection is the computational memory resources required. We evaluated this by measuring the RAM in gigabytes (GB) required for both classifiers. Figure 1d compares the RAM required for the single-threaded runs of Kraken/Bracken against the multi-threaded runs using QIIME. Notably, all Kraken/Bracken runs used less than 0.5 GB of RAM, which appears in the figure as zero GB. To provide more detail on RAM usage, Fig. 1e reports the RAM required by Kraken/Bracken in megabytes (MB) for all multi-threading options.

The resulting counts per genus for each of the human, ocean, and soil samples are listed in Supplemental Tables 1, 2, and 3, respectively. Figure 2 compares the true distribution of genera in each metataxonomic sample against the genus-level counts reported by Kraken 2, Bracken, and QIIME 2. For clarity, this figure shows the combined read counts across the V12, V34, V4, and V45 samples for each metagenome.

Table 1 16S rRNA databases used for the metataxonomic classifiers in this study

Database	Version	Release date	Sequences	Domains	Phyla	Classes	Orders	Family	Genera	Species
Greengenes	13_8	August 15, 2013	203,452	2	89	248	404	513	2102	2952
SILVA	132	December 13, 2017	695,171	5	228	514	1277	1531	9379	-
RDP	11.5	September 30, 2016	3,356,808	2	60	99	154	384	2466	-

For each of the most recently released versions of three 16S rRNA databases, this table describes the total number of sequences and the number of "traditional" nodes represented in their respective taxonomies. The Greengenes numbers refer to the 99% OTU database, and the SILVA values reflect the Ref NR 99 database

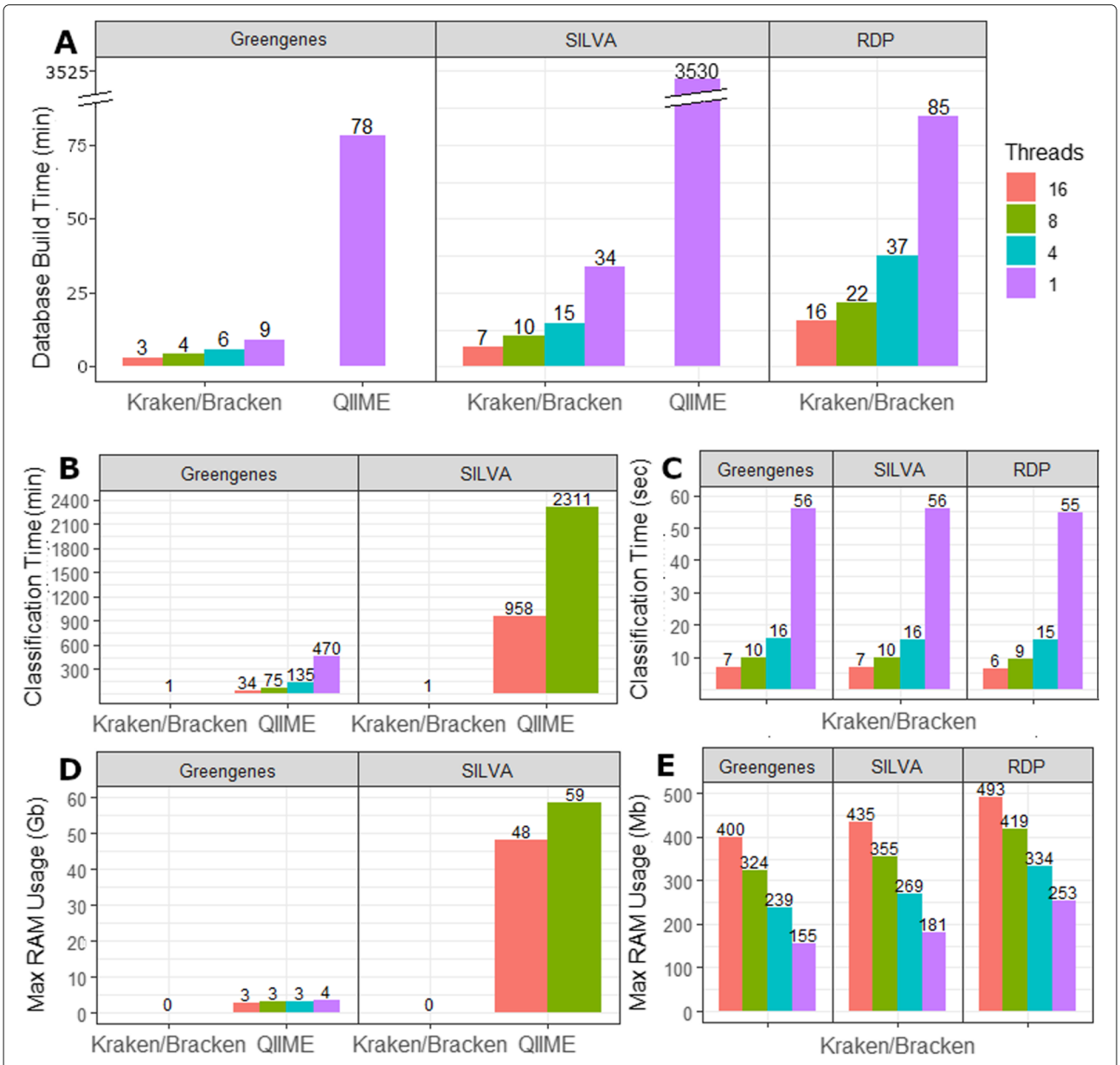

Fig. 1 Build and classification statistics. **a** Required time to build each database for Kraken/Bracken and QIIME. Kraken and Bracken allow for multi-threading while QIIME 2's q2-feature-classifier is single-threaded. **b** Average classification runtime in minutes for each database. Kraken/Bracken combined runtime is reported for only 1 thread as all runtimes are < 1 min and bars are too small to be visible at this scale. QIIME was only run using 16 and 8 threads for SILVA. **c** Classification runtime for Kraken and Bracken in seconds for all multi-threading options. **d** Computational memory usage (RAM) for QIIME and Kraken/Bracken, shown in gigabytes (Gb). Kraken/Bracken RAM requirements reported only for 1 thread as Kraken and Bracken require < 0.5Gb of RAM regardless of thread count. **e** Computational memory usage (RAM) for Kraken/Bracken shown in megabytes (Mb)

Table 2 Sample read counts

Read counts	V12	V34	V4	V45	Total
Human microbiome	186,689	189,972	193,787	192,319	762,767
Soil microbiome	196,254	193,564	196,226	194,325	780,369
Ocean microbiome	193,867	193,962	196,198	195,135	779,162

The read counts in each metagenome-primer sample. Each sample was generated as described in the Supplementary Methods

We used two different metrics to evaluate the genus distribution accuracy: mean absolute percentage error (*MAPE*) and Bray-Curtis dissimilarity (*BC*). Both error rates measure how different the predicted sample distribution is from the true genera counts. See the "Methods" section for details on how each error rate is calculated. Given these two metrics, we evaluate accuracy as $1 - MAPE$ and $1 - BC$. Figure 3a compares the accuracy of each tool when calculating the correct combined read counts at the genus level for each

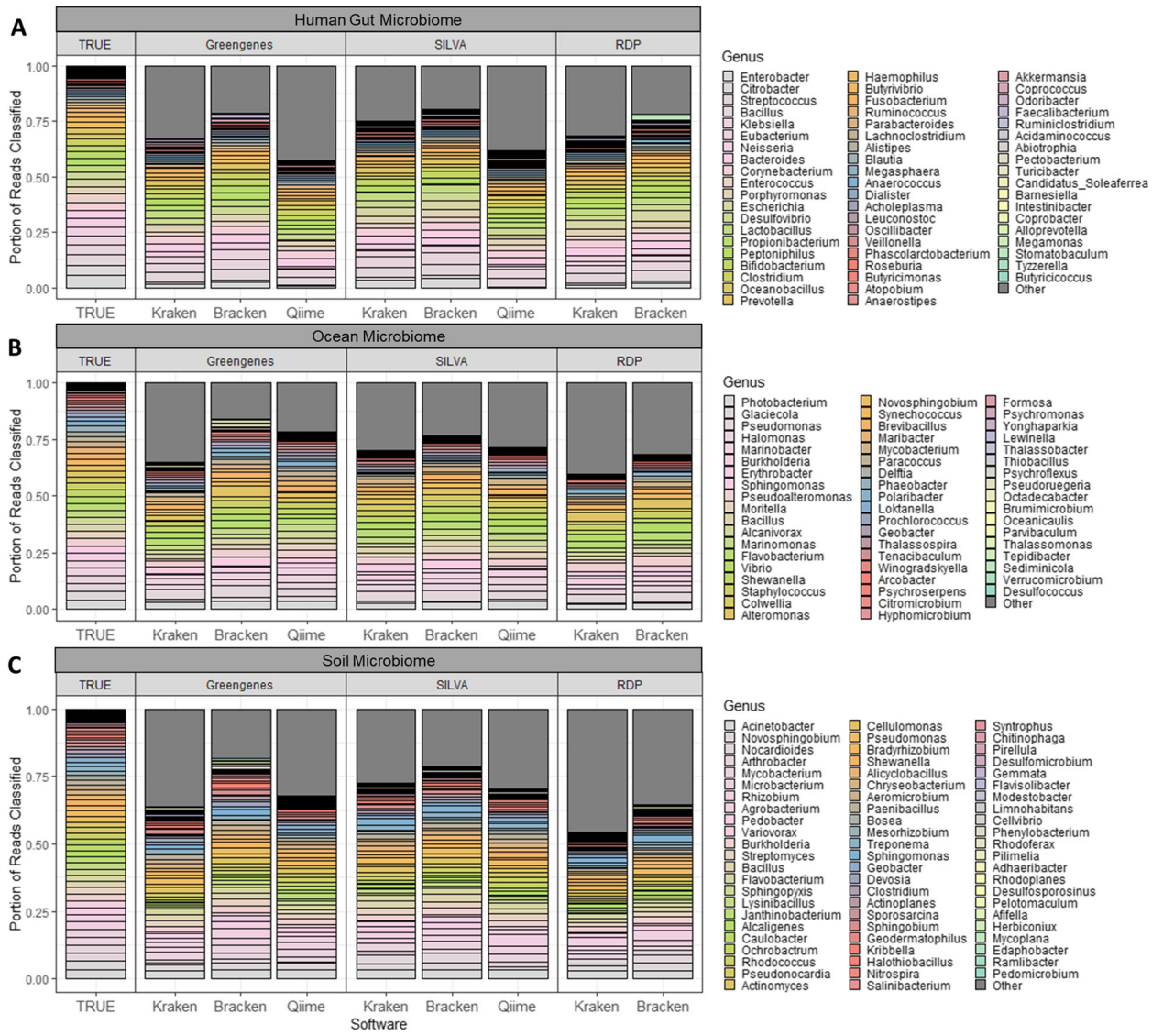

Fig. 2 Genera distribution for simulated microbiota. This plot compares the true genus relative abundances against those relative abundances estimated by Kraken, Bracken, and QIIME's q2-feature-classifier, for each of the three simulated microbiome samples (**a** human gut microbiome, **b** ocean microbiome, **c** soil microbiome). Only the correct genera are represented by different bars while read assignments to any incorrect taxon are included in "Other"

metagenome. For further insight into how the choice of 16S rRNA primer affects genus distribution accuracy, we evaluated the average *MAPE* and average *BC* across all 3 metagenome samples for each program/database. Figure 3b uses these averages to compare the accuracy between 16S rRNA primers. Supplemental Table 4 lists all *MAPE* and *BC* values for each combination of software/database/primer/metagenome.

While all tools tested provide general read counts per genus, Kraken is the only tool that directly assigns each read with a taxonomic label. Using this information, we can calculate Kraken's accuracy when classifying reads at major taxonomic levels in terms of sensitivity and precision. We measure precision by positive predictive value (PPV, see Supplemental Methods for more details). Figure 4 displays Kraken's average sensitivity and PPV for each database used (Fig. 4A) and for each 16S rRNA primer used in generating the samples (Fig. 4b).

Discussion

In this study, we evaluated three systems for classification and relative abundance estimation of 16S rRNA sequencing data sets: Kraken 2, Bracken, and QIIME 2. For Kraken and Bracken, we used three 16S rRNA databases: Greengenes, SILVA, and RDP, while for QIIME, we only evaluated Greengenes and SILVA. We then used these tools/databases to classify 12 samples generated by Almeida et al., which represent 3 simulated metagenomes (human gut, ocean, and soil) and 4 different 16S rRNA primers (V12, V34, V4, and V45). In total, we collected 36 different results using Kraken/Bracken and 24 different results using QIIME.

Database building time

For all systems compared here, database build time is a function of the number of sequences in the database. Because 16S Greengenes is the smallest database (with 200,000 sequences) and 16S RDP is the largest (with 3.4 million sequences), generation of database files is fastest with Greengenes and slowest with RDP.

When comparing single-threaded Kraken/Bracken against QIIME's q2-feature-classifier, Kraken and Bracken combined require far less build time. For the smallest 16S rRNA database, Greengenes, QIIME required more than an hour to generate the naïve Bayes classifier (Fig. 1a). By comparison, single-threaded Kraken and Bracken combined required less than 10 min to create the database files. For 16S SILVA, with nearly 700,000 sequences, QIIME 2 required more than 58 h for classifier generation while the single-threaded Kraken/Bracken required only $\sim$ 30 min. We additionally note that the largest 16S rRNA database, RDP, required a little more than an hour for single-threaded Kraken 2 and Bracken to create the database files. As mentioned above, the RDP database is incompatible with QIIME 2. The multi-threaded nature of Kraken 2 and Bracken further accelerate the database building process, with 4 threads halving the required build time (Fig. 1a).

Classification time/memory requirements

As observed by Almeida et al., QIIME 2's q2-feature-classifier requires more computational resources than other methods during classification. With the use of 16 CPU threads, QIIME required 35 min on average to classify the human, ocean, and soil metataxonomic samples using the Greengenes database (Fig. 1b). The QIIME's SILVA classifier required 16 h on average. By comparison, single-threaded Kraken 2 and Bracken required on average 1 min per metataxonomic sample. This runtime decreases from 1 min to 15, 10, and 6 s for 4, 8, and 16 threads respectively (Fig. 1c). The runtime of Kraken 2 and Bracken was nearly the same for all three databases. Thus, Kraken or Braken is at least 350 times faster (6 s vs. 35 min) than QIIME 2 when run with 16 parallel threads.

The amount of computer memory (RAM) required by each system also varied widely (Fig. 1d). For all three databases, single-threaded Kraken required $<$ 260 MB of RAM. However, the single-threaded QIIME Greengenes classifier required 3.6 GB of RAM. Increasing the number of threads for Kraken also increases the total RAM used, with 16 threads using 400–500 MB of RAM for each of the Kraken databases (Fig. 1e). However, for QIIME, increasing the number of threads decreased the total RAM: the QIIME Greengenes classifier with 16 threads used $\sim$ 2.7 GB, and the QIIME SILVA classifier with 16 threads used 48 GB of RAM (Fig. 1d).

Accuracy of relative abundance estimation

Finally, we compared the accuracy of all three tools based on their ability to recreate the true genus distribution of the simulated samples (Fig. 2). We quantified the accuracy of these distributions using both MAPE and Bray-Curtis dissimilarity (Supplemental Table 4).

In all cases, Bracken performed better than Kraken 2, which was expected because Kraken is a classification tool, not an abundance estimation system. Kraken classifies reads at any level in the taxonomy, which means that some reads might be assigned to a higher level genus, e.g., any read that has equally good matches to two genera will be assigned to the family containing them. For the simulated datasets in this study, Kraken assigned from 7–30% of the reads to levels above genus. These reads are not incorrectly classified, but the result is that Kraken underestimates the abundances of their genera. By contrast, Bracken is designed to use Kraken's classification data to estimate all read counts at the genus level, thereby improving on Kraken's genus-level distribution.

On average, Bracken performed the best, having the lowest average error rates across all three 16S rRNA databases (Supplemental Table 4). Bracken also had the lowest error rate for 8/9 combinations of samples and databases. The only sample where QIIME 2 had a lower error rate than Bracken was in the classification of the ocean samples against the 16S Greengenes database (Fig. 3a). However, QIIME 2 had the highest error rate when classifying the human sample against Greengenes or SILVA, regardless of whether measured by MAPE or Bray-Curtis dissimilarity.

In analyzing the trends across the databases using both MAPE and Bray-Curtis, Bracken performed the best using the 16S SILVA database and performed the worst using the 16S RDP database. 16S RDP yielded on average 0.391 MAPE and 0.221 BC Index while 16S SILVA only yielded a 0.286 MAPE and a 0.153 BC Index. 16S Greengenes with Bracken had an average of 0.313 MAPE and a 0.165 BC Index. Although QIIME 2 was not tested on 16S RDP, QIIME 2 yielded the same trends when comparing 16S Greengenes and SILVA, with 16S SILVA outperforming 16S Greengenes in almost all cases.

In addition to evaluating the different tools, we also evaluated the accuracy of each of the primer sets (V12, V34, V4, and V45) that were used by Almeida et al.. Figure 3b shows the average accuracy of each primer set across all 3 metagenomes for a given software/database pairing. For both Greengenes and SILVA, the samples generated using V34 and V12 performed slightly better. However, for RDP, the difference in accuracy between primer samples is further magnified. When classifying with the RDP database, both Kraken and Bracken had significantly better results for the V12 and V34 samples (Fig. 3b).

Per-read classification accuracy

Kraken is the only program of the three tested here that provide per-read assignments by default, allowing us to compute the read-level accuracy of its taxonomy assignments. Per-read accuracy is somewhat dependent on the reference database, but highly dependent on the 16S

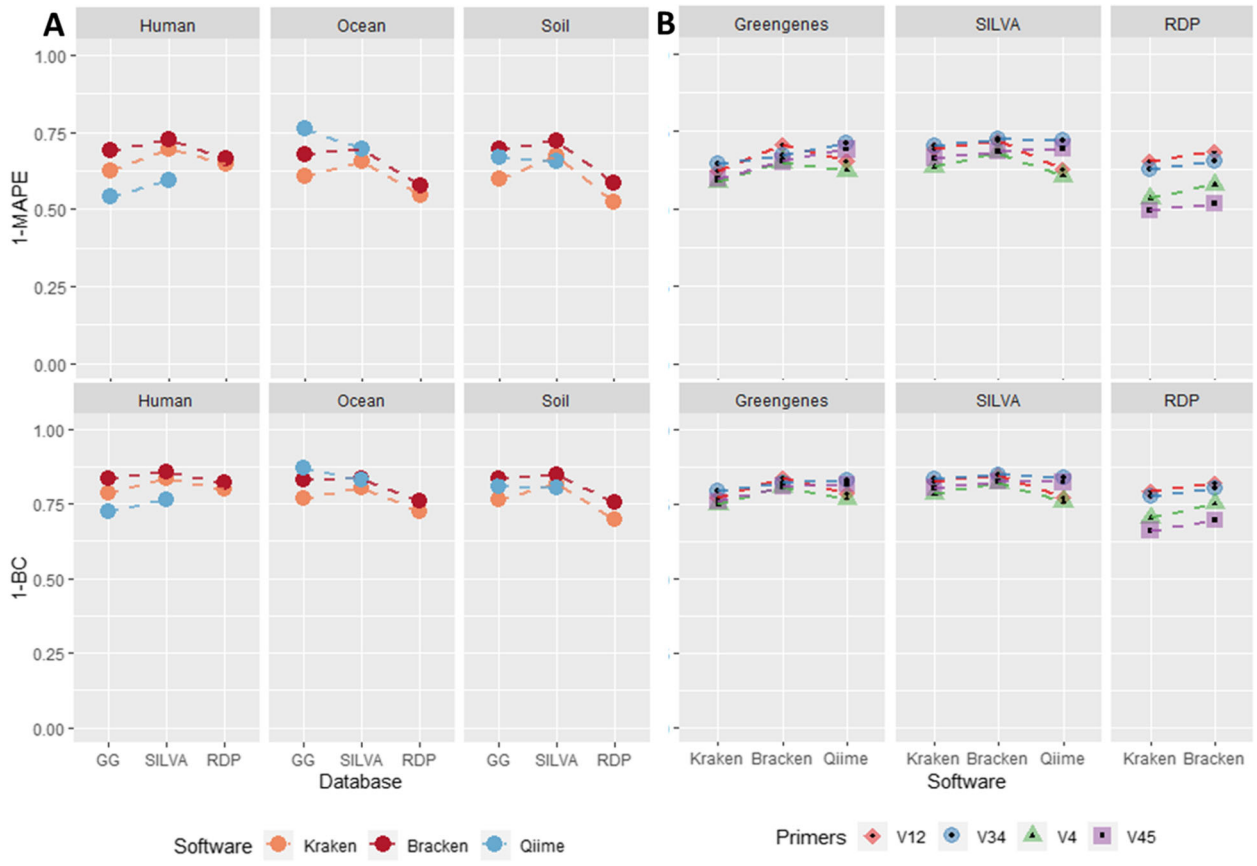

Fig. 3 MAPE and Bray-Curtis dissimilarity. This plot evaluates classification accuracy by using the inverse of two error metrics: mean absolute proportion error (*MAPE*) and Bray-Curtis dissimilarity (*BC*). **a** Comparison of the accuracy of Kraken, Bracken, and Qiime's q2-feature-classifier when predicting the genus read counts across all samples for given metagenome/database. **b** Comparison of the accuracy between the individual primers averaged across all 3 metagenomes for a given software/database. The top plots calculate accuracy as 1 — *MAPE* while the bottom plots evaluate 1 — *BC*

rRNA primer set (Fig. 4b). In particular, Kraken had three times higher sensitivity (60%) and PPV (65%) when classifying reads generated using V12 primers versus those generated from V45 primers (20% and 21%).

As expected, sensitivity and precision increased with taxonomic level, with class and phylum sensitivity and precision exceeding 0.95 for all sample sets and all databases. Supplemental Table 6 contains exact numbers for sensitivity and precision for each dataset and database.

Taxonomy inconsistencies

In our experiments, we observed that the accuracy of 16S rRNA analysis is highly dependent on the choice of 16S rRNA database, a phenomenon well known to the 16S rRNA community [13, 25]. The 170 distinct genera present in our human, ocean, and soil metagenomes were selected from the NCBI taxonomy, but none of the three 16S rRNA database taxonomies contains precisely the same genera. Each 16S rRNA database is independently curated from different reference sets, resulting in substantial differences among the taxonomies [41]. Among the 170 unique genera uses here, 22 are missing from Greengenes, 19 have different names or are mapped to multiple genera in RDP, and 16 have different names in Silva (see Supplemental Table 5). For example, Agrobacterium, Burkholderia, and Rhizobium are not unique genera in the 16S SILVA taxonomy, but are combined into a single "Allorhizobium-Neorhizbium-Pararhizobium-Rhizobium" genus. Escherichia and Shigella are also combined into the "Escherichia-Shigella" genus in 16S SILVA. The Clostridium sequences in 16S SILVA are split between 19 different genera, each with the prefix of "Clostridium sensu stricto" followed by a number 1–19 [42].

Conclusion

Although each of the 16S rRNA databases represents a large number of bacterial organisms, the accuracy of metataxonomic classifiers varied substantially among them. In our experiments, 16S SILVA provided the lowest error rates and highest per-read accuracy regardless of the software used in classification. Across all databases, Kraken 2 and Bracken outperformed QIIME 2's q2-feature-classifier in terms of computational

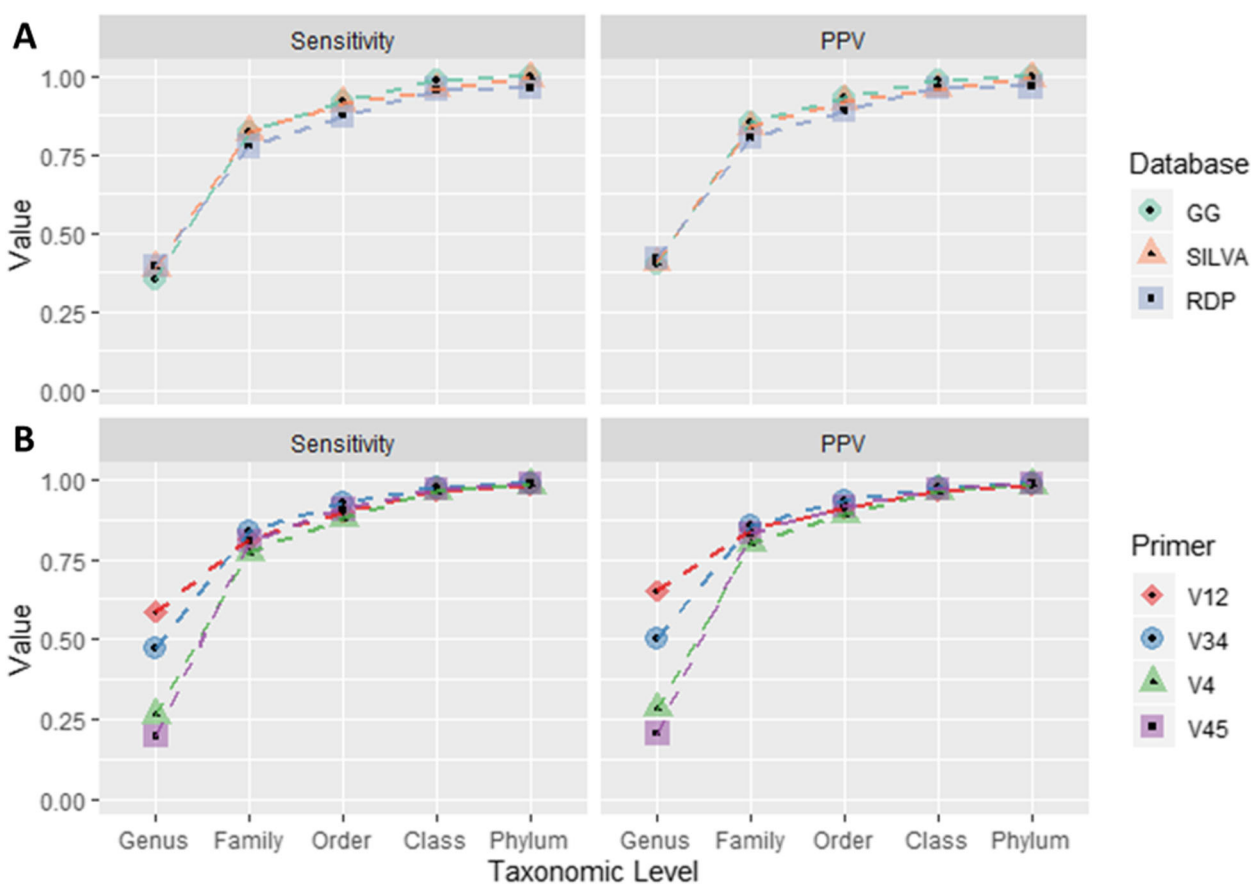

Fig. 4 Kraken per-read accuracy. As Kraken is the only tool tested that provides per-read taxonomy assignments, we evaluate the sensitivity and precision (*PPV*) of Kraken 2's taxonomy assignments at each major taxonomic level

requirements, runtime, and accuracy. Single-threaded Kraken/Bracken was nearly 8x faster than QIIME 2 at building the 16S Greengenes database and 100x faster at building a 16S SILVA database. Kraken and Bracken also allow for multi-threaded database building, which allows any 16S rRNA database to be built in less than 20 min. For classification, Kraken/Bracken used 20 times less RAM, performed 300 times faster, and achieved better genus-level resolution than QIIME 2.

Methods

Almeida simulated data

QIIME 2, Kraken 2, and Bracken were evaluated using the A500 synthetic microbiome samples generated by Almeida et al. and available at ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/taxon_benchmarking/. The A500 set contains 12 samples representing three different microbial environments: the human gut, ocean, and soil. For each of these environments, genomic sequences for their most abundant genera were extracted and randomly sampled. These human gut, ocean, and soil genomes then were sub-sampled four times to simulate 16S rRNA profiling using four different primer sets, generating 200,000,250-bp paired-end reads per primer sequence. The sub-sampling introduced a 2% random mutation to each sequencing read. Almeida et al. then performed pre-processing and quality control to filter sequences with ambiguous base calls, as is suggested for QIIME 2 analysis ([30, 43]). With three microbial environments and four primer sets, Almeida et al. thereby generated 12 sets of synthetic communities for testing. Information about the software and primers used in dataset generation is further described in the "Methods" section of Almeida et al. [34].

Software and databases

The software packages tested are Kraken 2 (downloaded on 2020/03/05), Bracken v2.5, and QIIME 2 v2017.11. Kraken and Bracken database files were generated for Greengenes 13_8, SILVA 132, and RDP 11.5 database releases. QIIME 2 database files were generated for Greengenes 13_8 and SILVA 132.

Error rate calculations

For evaluating the accuracy of Kraken 2, Bracken, and QIIME 2, we calculated two different error metrics which compare the true genera distributions against those reported by each program. The first error metric is a modified mean absolute proportion error (MAPE) which compares the difference between the true read counts (T_g) for a given genus and the measured read counts (A_g) for that same genus.

$$MAPE = \sum_{g=1}^{n} \frac{T_g}{\sum_{g=1}^{n} T_g} \times \frac{|A_g - T_g|}{T_g} \tag{1}$$

Each difference is calculated as a fraction of the true counts and then weighted by the fraction of the total sample. n is the total number of true genera in the sample.

The second metric, Bray-Curtis dissimilarity [44], is a similar measurement of the dissimilarity between the true genera distribution and the measured genera distribution. The formula for Bray-Curtis dissimilarity is:

$$BC_ij = 1 - \frac{2C_ij}{S_i + S_j} \tag{2}$$

where C_ij is the sum of lesser reads for genera in common and $S_i = S_j$ is the total number of reads. In other words, for every true genus g in the sample, if $T_g < A_g, C_ij = C_ij + T_g$. Otherwise, if $T_g > A_g, C_ij = C_ij + A_g$.

MAPE and *BC* values both fall between 0 and 1, where larger values indicate a greater difference between samples and smaller values indicate a greater similarity.

Sensitivity and precision (PPV) calculations

As Kraken 2 provides taxonomic assignments for every read, we can use the true taxonomic tree of each read to calculate sensitivity and precision at all taxonomic levels. For this explanation, we describe our calculations of sensitivity and precision for the genus level. First, we calculate true positive (*TP*), vague positive (*VP*), false positive (*FP*), and false negative (*FN*) read counts. We define *TP* read counts as the number of reads correctly classified at the genus level. This includes reads that are classified as any species within the true genus. Vague positive (*VP*) reads account for the possibility that a read is classified as any ancestor of the true taxon. Therefore, *VP* reads include all *TP* reads and all reads assigned to ancestor taxa of the true genus. *FN* reads are all classified reads that are not *VP* reads. This thereby includes reads classified at any taxa not within the direct lineage of the true genera. Finally, we define *FN* as the number of unclassified reads. Notably, in all experiments, Kraken 2 did not label any read as unclassified (*FN* = 0).

From these values, we define sensitivity and precision (measured by positive predictive value, *PPV*) using the following two equations:

$$\begin{aligned} Sensitivity &= \frac{TP}{TP + VP + FN + FP} \\ &= \frac{TP}{TP + VP + FP} \end{aligned} \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40168-020-00900-2.

Additional file 1: This file contains all command lines used for testing Kraken 2, Bracken, and QIIME 2 and with building each of the 16S rRNA databases. Additionally, this file contains a short discussion of the multi-threading behavior of QIIME 2.

Additional file 2: Human Genus Read Counts – True v Kraken v Bracken v QIIME.

Additional file 3: Ocean Genus Read Counts – True v Kraken v Bracken v QIIME.

Additional file 4: Soil Genus Read Counts – True v Kraken v Bracken v QIIME.

Additional file 5: MAPE and Bray-Curtis Dissimilarities.

Additional file 6: 16S rRNA Database Inconsistencies.

Additional file 7: Kraken 2 Per Read Sensitivity and Precision.

Abbreviations

GG: Greengenes; TP: True positive; TN: True negative; FP: False positive; FN: False negative; MAPE: Mean absolute percentage error; BC: Bray-Curtis dissimilarity; PPV: Positive predictive value

Acknowledgements

We thank Alexandre Almeida for his assistance with acquiring the data for analysis.

Authors' contributions

JL conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, performed the computation work, and wrote the paper. SLS conceived and designed the experiments, analyzed the data, and wrote the paper. Both authors read and approved the final manuscript.

Funding

This work was supported in part by NIH under grants R01-HG006677 and R35-GM130151 and by NSF under grant IOS-1744309.

Availability of data and materials

The datasets analyzed during this study are provided in by Almeida et al., http://dx.doi.org/10.5524/100448. Data generated during this study is included in this published article and its supplementary information files.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. [3]Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD, USA.

Received: 6 April 2020 Accepted: 24 July 2020
Published online: 28 August 2020

References

1. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, Lewis BJ, Stahl D. Conservation of primary structure in 16S ribosomal RNA. Nature. 1975;254(5495):83–86.
2. Woese CR. Bacterial evolution. Microbiol Rev. 1987;51(2):221–71.
3. Gilbert JA, Jansson JK, Knight R. The earth microbiome project: successes and aspirations. BMC Biol. 2014;12:69.
4. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci USA. 2012;109(52):21390–5.
5. Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. 2010;4(10):1340–51.
6. Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, Delgardio J, Norton N, Hazen TC, Huettel M. Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill. Appl Environ Microbiol. 2011;77(22):7962–74.
7. Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, Fernandez-Guerra A, Jeanthon C, Rahav E, Ullrich M, Wichels A, Gerdts G, Polymenakou P, Kotoulas G, Siam R, Abdallah RZ, Sonnenschein EC, Cariou T, O'Gara F, Jackson S, Orlic S, Steinke M, Busch J, Duarte B, Caçador I, Canning-Clode J, Bobrova O, Marteinsson V, Reynisson E, Loureiro CM, Luna GM, Quero GM, Löscher CR, Kremp A, DeLorenzo ME, Øvreås L, Tolman J, LaRoche J, Penna A, Frischer M, Davis T, Katherine B, Meyer CP, Ramos S, Magalhães C, Jude-Lemeilleur F, Aguirre-Macedo ML, Wang S, Poulton N, Jones S, Collin R, Fuhrman JA, Conan P, Alonso C, Stambler N, Goodwin K, Yakimov MM, Baltar F, Bodrossy L, Van De Kamp J, Frampton DM, Ostrowski M, Van Ruth P, Malthouse P, Claus S, Deneudt K, Mortelmans J, Pitois S, Wallom D, Salter I, Costa R, Schroeder DC, Kandil MM, Amaral V, Biancalana F, Santana R, Pedrotti ML, Yoshida T, Ogata H, Ingleton T, Munnik K, Rodriguez-Ezpeleta N, Berteaux-Lecellier V, Wecker P, Cancio I, Vaulot D, Bienhold C, Ghazal H, Chaouni B, Essayeh S, Ettamimi S, Zaid EH, Boukhatem N, Bouali A, Chahboune R, Barrijal S, Timinouni M, El Otmani F, Bennani M, Mea M, Todorova N, Karamfilov V, Ten Hoopen P, Cochrane G, L'Haridon S, Bizsel KC, Vezzi A, Lauro FM, Martin P, Jensen RM, Hinks J, Gebbels S, Rosselli R, De Pascale F, Schiavon R, Dos Santos A, Villar E, Pesant S, Cataletto B, Malfatti F, Edirisinghe R, Silveira JAH, Barbier M, Turk V, Tinta T, Fuller WJ, Salihoglu I, Serakinci N, Ergoren MC, Bresnan E, Iriberri J, Nyhus PAF, Bente E, Karlsen HE, Golyshin PN, Gasol JM, Moncheva S, Dzhembekova N, Johnson Z, Sinigalliano CD, Gidley ML, Zingone A, Danovaro R, Tsiamis G, Clark MS, Costa AC, El Bour M, Martins AM, Collins RE, Ducluzeau A-L, Martinez J, Costello MJ, Amaral-Zettler LA, Gilbert JA, Davies N, Field D, Glöckner FO. The ocean sampling day consortium. Gigascience. 2015;4:27.
8. Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E, Schulze-Lefert P. Structure and functions of the bacterial microbiota of plants. Annu Rev Plant Biol. 2013;64:807–38.
9. Hardoim PR, van Overbeek LS, Berg G, Pirttilä AM, Compant S, Campisano A, Döring M, Sessitsch A. The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. Microbiol Mol Biol Rev. 2015;79(3):293–320.
10. Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL, Buckler ES, Ley RE. Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci USA. 2013;110(16):6548–53.
11. Patel JB. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Mol. Diagn. 2001;6(4):313–21.
12. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761–4.
13. Clarridge 3rd JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004;17(4):840–62.
14. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen AM, Peet A, Tillmann V, Pöhö P, Mattila I, Lähdesmäki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Orešič M, Huttenhower C, Knip M, DIABIMMUNE Study Group, Xavier RJ. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17(2):260–73.
15. Emery DC, Shoemark DK, Batstone TE, Waterfall CM, Coghill JA, Cerajewska TL, Davies M, West NX, Allen SJ. 16S rRNA next generation sequencing analysis shows bacteria in Alzheimer's post-mortem brain. Front Aging Neurosci. 2017;9:195.
16. Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, Benyamin FW, Lei YM, Jabri B, Alegre M-L, Chang EB, Gajewski TF.

Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. Science. 2015;350(6264):1084–9.
17. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155(7):1451–63.
18. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
19. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. Nature. 2019;569(7758):641–8.
20. Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, Koren O, Fierer N, Kelley ST, Ley RE, Gordon JI, Knight R. Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. 2010;11(5):210.
21. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
22. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72.
23. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):590–6.
24. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42(Database issue):633–42.
25. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
27. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.
28. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28(24):3211–7.
29. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson 2nd MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7.
30. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É.. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–30.
32. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:2584.
33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
34. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience. 2018;7(5):giy054.
35. Matias Rodrigues JF, Schmidt TSB, Tackmann J, von Mering C. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics. 2017;33(23):3808–10.
36. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
37. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):46.
38. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3:104.
39. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(257):762302.
40. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:19233.
41. Balvočiūtė M, Huson DH. SILVA, RDP, greengenes, NCBI and OTT - how do these taxonomies compare?. BMC Genomics. 2017;18(Suppl 2):114.
42. Knight DR, Elliott B, Chang BJ, Perkins TT, Riley TV. Diversity and evolution in the genome of clostridium difficile. Clin Microbiol Rev. 2015;28(3):721–41.
43. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
44. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr. 1957;27(4):325–49.

Publisher's Note