scientific reports

Check for updates

OPEN

Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data

Aubrey R. Odom[1,2,7], Tyler Faits[1,2,7], Eduardo Castro-Nallar[3,4], Keith A. Crandall[5] & W. Evan Johnson[6]✉

Most experiments studying bacterial microbiomes rely on the PCR amplification of all or part of the gene for the 16S rRNA subunit, which serves as a biomarker for identifying and quantifying the various taxa present in a microbiome sample. Several computational methods exist for analyzing 16S amplicon sequencing. However, the most-used bioinformatics tools cannot produce high quality genus-level or species-level taxonomic calls and may underestimate the potential accuracy of these calls. We used 16S sequencing data from mock bacterial communities to evaluate the sensitivity and specificity of several bioinformatics pipelines and genomic reference libraries used for microbiome analyses, concentrating on measuring the accuracy of species-level taxonomic assignments of 16S amplicon reads. We evaluated the tools DADA2, QIIME 2, Mothur, PathoScope 2, and Kraken 2 in conjunction with reference libraries from Greengenes, SILVA, Kraken 2, and RefSeq. Profiling tools were compared using publicly available mock community data from several sources, comprising 136 samples with varied species richness and evenness, several different amplified regions within the 16S rRNA gene, and both DNA spike-ins and cDNA from collections of plated cells. PathoScope 2 and Kraken 2, both tools designed for whole-genome metagenomics, outperformed DADA2, QIIME 2 using the DADA2 plugin, and Mothur, which are theoretically specialized for 16S analyses. Evaluations of reference libraries identified the SILVA and RefSeq/Kraken 2 Standard libraries as superior in accuracy compared to Greengenes. These findings support PathoScope and Kraken 2 as fully capable, competitive options for genus- and species-level 16S amplicon sequencing data analysis, whole genome sequencing, and metagenomics data tools.

High-throughput sequencing has greatly accelerated the study of microbiomics, that is, the scientific field focused on studying the composition, diversity, and function of microbial communities and their interactions with their hosts or environments[1]. Characterizing the composition of microbial samples commonly relies on the amplification of 16S ribosomal subunit sequences, a ubiquitous gene with highly conserved regions. The subunit simplifies efforts to isolate and amplify 16S rRNA with established PCR primers and hypervariable regions to establish identity and phylogeny. 16S rRNA and rDNA sequencing can be used to identify known prokaryotic species and act as a proxy to quantify the relative abundances of operational taxonomic units (OTUs) within microbiome samples.

Methods for taxonomic profiling of ribosomal RNA gene sequences enable sample OTU identification by classifying rRNA sequences into taxonomic groups. While considerable accuracy in species-level identification is attainable with available tools[2], current profiling software for 16S amplicon sequencing data hesitates to identify down to the species level. Instead, they cluster reads based on sequence similarity to assign genus or higher-level identifications to increase specificity and sensitivity, or they directly use error-filtered sequences for taxonomic classification[3,4]. As the capabilities of modern sequencing platforms increase, and as bacterial reference genome

[1]Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA. [2]Bioinformatics Program, Boston University, Boston, MA, USA. [3]Departamento de Microbiología, Facultad de Ciencias de la Salud, Universidad de Talca, Campus Talca, Avda. Lircay S/N, Talca, Chile. [4]Centro de Ecología Integrativa, Universidad de Talca, Campus Talca, Avda. Lircay S/N, Talca, Chile. [5]Department of Biostatistics & Bioinformatics, Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA. [6]Division of Infectious Disease, Center for Data Science, Rutgers University – New Jersey Medical School, Newark, NJ, USA. [7]These authors contributed equally: Aubrey R. Odom and Tyler Faits. ✉email: w.evan.johnson@rutgers.edu

databases expand and improve, more potential arises for achieving enhanced 16S analysis performance with alternative methods more commonly applied in whole genome metagenomics.

The most common software packages currently employed in analysis of 16S amplicon sequencing data are DADA2[4], QIIME 2[5], its predecessor, QIIME 2[6], and Mothur[7]. QIIME 2 and Mothur were both originally developed shortly after the invention of next-generation sequencing and, along with QIIME 2, essentially follow the same workflow: reads are typically clustered de novo based on sequence similarity into operational taxonomic units (OTUs) or denoised OTUs (many refer to these as amplicon sequence variants or ASVs) depending on whether complete sequence identity is desired for clustering. The initial clustering step serves to 1) improve computational efficiency by limiting the number of sequences needing alignment to a large set of reference genomes and 2) accommodate the low levels of genetic variation present within a given bacterial strain, thereby mitigating sequencing errors. For nearly a decade, the cutoff for OTU inclusion was 97% sequence identity[8,9], but current cutoff recommendations are now around 99–100% sequence identity[3,10], typically after some form of denoising or other correction for sequencing errors[4,11].

An alternative to OTU clustering includes directly aligning reads against a reference genome library, as is done by PathoScope 2.0[12]. PathoScope employs a Bayesian mixed modeling framework to reassign ambiguously aligned reads, dampening potential sequencing errors and minor genetic variation[13,14]. As another alternative, Kraken 2 performs alignment-free *k*-mer searches against a reference genome library[15] and makes taxonomic assignments to each read based on the cumulative number of *k*-mer matches across an entire read against each taxonomic node in its reference library. By bypassing a sequence clustering step, PathoScope and Kraken 2 individually avoid the potential pitfalls inherent in OTU generation and denoising errors[16,17], although remain susceptible to sequencing errors. While DADA2, QIIME 2, Mothur, Greengenes, and SILVA are all tools designed to address the specific needs of 16S amplicon sequencing, improvements in sequencing technologies, expanding bacterial reference genome databases, and increased availability and affordability of computational resources have collectively made many of the specific issues addressed by these tools irrelevant. Meanwhile, the increased flexibility and power of a tool such as PathoScope may yield augmented results despite being computationally intensive and designed to fulfill a more general metagenomics purpose[18,19].

These profiling methods all rely heavily on the quality of the reference library used, as has been shown in previous benchmarking studies[20–23]. The most commonly used reference databases for 16S amplicon analyses are Greengenes[24], SILVA[25], and the Ribosomal Database Project (RDP)[26]. Each database exclusively contains 16S rRNA gene sequences and offers taxonomic information for each reference sequence. SILVA is well maintained and releases updates regularly, although as of this writing, the most recent update is SILVA 138.1 (released on August 27, 2020). Meanwhile, Greengenes has been stagnant for years; its most recent update at the time of submission was Greengenes 13_8, released in August 2013. As a result, Greengenes lacks several essential bacteria, including *Dolosigranulum* species[27], implicated as playing a protective role in preventing disease in human airways[28,29]. It is worth noting that a new version, Greengenes2[30], was made available in 2022 amidst the review process for this paper, and was subsequently not included in this paper's analyses. Although QIIME 2 and Mothur are compatible with any reference genome library, QIIME 2 uses Greengenes by default, and Mothur's documentation (as accessed on May 17, 2022) recommends SILVA. DADA2 maintains reference databases for SILVA, RDP, and Greengenes, with the flexibility to create custom databases. Kraken 2 has its own curated "Standard" bacterial library, with a taxonomic tree based on NCBI's taxonomy database by default[31], and has also released Kraken 2-compatible formatted versions of Greengenes, SILVA, and RDP. The current PathoScope reference library recommendation is to download the complete RefSeq representative genome database[32], a collection of curated high-quality bacterial genomes and assemblies. RefSeq is constantly updated, and as such, results of any analysis using RefSeq as a reference library may vary according to the date of the library download.

Given these considerations, we systematically benchmarked several current community profiling tools and reference libraries created for both metagenomic and 16S analysis. We evaluated the tools QIIME 2, Mothur, PathoScope 2, and Kraken 2 in conjunction with reference libraries from Greengenes, SILVA, Kraken 2, and RefSeq. Using several publicly available 16S sequencing datasets of synthetic mock communities, we specifically analyzed genus- and species-level performance across pairs of profilers and libraries. We tested 136 samples comprising varying species richness and evenness, several different amplified regions within the 16S rRNA gene, and both DNA spike-ins and cDNA from collections of plated cells. Our evaluative comparisons utilized a combination of diversity and accuracy-based measures to determine what methods and tools provided the best performance in profiling 16S amplicon sequencing datasets.

Methods

Publicly available mock community sequencing datasets. 136 mock community sequencing samples were collected in total from four publicly available sequencing datasets and analyzed in our evaluation. 69 samples are from Lluch et al.[33]; 33 samples are from Kozich et al.[34]; 29 samples are from Fouhy et al.[35]; and 5 samples are from Karstens[36]. These datasets are hereafter referred to as the Lluch, Kozich, Fouhy, and Karstens samples. The species compositions for each set are delineated in Table S1. The Lluch samples include a variety of community compositions, ranging from monoculture samples composed of only a single species to others with 20 species at staggered concentrations. Collectively, 34 species appear across the aggregate of Lluch samples. While the Lluch samples' taxonomic profiles are diverse, all 69 samples were produced using a single unified DNA extraction, amplification, and sequencing protocol that yielded Illumina MiSeq paired-end reads of the V4–V5 region of the 16S rRNA gene. The Kozich samples each comprise three sequencing replicates of 11 different preparations of BEI's mock community B (HM-278D), encompassing 21 species. For the Kozich samples, three PCR primer pairs were used to amplify three distinct portions of the 16S rRNA gene (the V3, V4, and V4-V5 ranges), making the sequencing data for these samples more complex than for those samples from the

other datasets. The Fouhy samples are each a unique combination of either BEI mock community B (16S DNA spike-ins) or BEI mock community C (cultured cells), prepared using one of three library prep protocols, amplified with PCR primers for either the V1-V2 or the V4-V5 region of the 16S rRNA gene, and sequenced either on an Illumina MiSeq machine or a Thermo Fisher Ion Torrent. Finally, the 5 Karstens samples originate from a single custom mock DNA library of 8 species, with the V4 region amplified and sequenced on an Illumina MiSeq device.

16S amplicon sequencing analysis pipelines. We evaluated five analysis pipelines applied to the 136 mock community samples: DADA2, QIIME 2, Mothur, PathoScope, and Kraken 2.

For the standalone implementation of DADA2, all samples were filtered and trimmed, with errors learned on forward and reverse reads. Learned errors were used to conduct inference on predicted error presence across all reads as a denoising measure. Paired reads were merged and chimeras were removed, with taxonomy assigned down to the species level. When filtering and trimming, for most samples we used the parameters maxN = 0, maxEE = c(3, 3), truncQ = 2, rm.phix = TRUE, and tlength = 0. For unpaired reads, we set maxEE = 3. Several Kozich samples suffered from extreme quality degradation at read ends; to correct for this, we set tlength = c(240, 200). Finally, when running the Fouhy Ion Torrent samples, we set the DADA2 function parameters HOMOPOLYMER_GAP_PENALTY = − 1 and BAND_SIZE = 32. For the filter and trim step, we also set trimLeft = 15. These settings were based on recommendations for processing Ion Torrent data in the DADA2 FAQ.

For all QIIME 2 analyses, we used the DADA2 plugin to cluster sequences and construct feature tables. We decided to test the DADA2 plugin alongside its standalone package due to the broad user base of QIIME 2. However, the standalone implementation uses 100% sequence identity over Mothur and QIIME 2 (with 97%) and uses exact sequence matching rather than a k-mer based method (as in the QIIME 2 q2-feature-classifier). All mock datasets could be run with paired-end reads besides the Fouhy datasets. In most cases, DADA2 did not require truncation of paired-end sequences and only the initial 6 bp were trimmed from each read. However, quality scores at the end of nine samples from the Kozich dataset were universally low enough (cutoff of median quality score < 20) to require truncation to 240 bp for forward reads and 200 bp for reverse reads for Kozich samples. Taxonomy was assigned using custom naïve Bayes classifiers constructed for each set of mock community samples based on their amplified 16S region. The QIIME 2 artifact output files were converted into BIOM format and subsequently into tab-delimited text format for downstream analyses and pipeline comparisons.

For Mothur analyses, all recommended procedures were followed according to Mothur documentation where possible. For paired-end sequences, the native make.contigs() function was used to join reads. In the pre.cluster() step of Mothur analysis, the "diffs" parameter (the number of mismatches allowed between a cluster's representative sequence and each member sequence) was set to 2 for joined sequencing reads shorter than 250 bp, 3 for joined reads of length 250–349 bp, and 4 for longer joined reads. For cluster.split(), we set the "taxlevel" parameter to 4, with a "cutoff" of 0.03.

For PathoScope 2.0 analyses, Bowtie2 alignment parameters were set to "–local -R 2 -N 0 -L 25 -i S,1,0.75 -k 10 –score-min L,100,1.28." These values were optimized for 16S sequencing reads, requiring higher similarity to a reference genome to be considered a hit than the default settings due to the highly conserved nature of portions of the 16S rRNA gene. Phylogeny for each taxon was inferred from the NCBI taxon id (ti) for each reference genome using the entrez_fetch() function from the R package *rentrez*.

For Kraken 2 analyses, Kraken 2 taxonomic reports were created for each sample. These were parsed into a taxon/feature counts matrix that included the complete phylogeny for each identified taxon as reported by Kraken 2.

Bacterial genomic and 16S reference libraries. We used five bacterial sequence reference databases in conjunction with the aforementioned pipelines: Greengenes 13_8, SILVA 138, two versions of RefSeq's representative genomes, and the Kraken 2 Standard library (downloaded on August 20, 2020). According to the Kraken 2 manual, the Kraken 2 Standard library is compiled using the RefSeq database, so it could be considered analogous to the RefSeq2020 library. The RefSeq libraries were downloaded on November 2, 2018, and June 23, 2020; these are denoted as "RefSeq2018" and "RefSeq2020." Greengenes and SILVA are specifically 16S reference databases as they include only sequences for the bacterial 16S rRNA gene. RefSeq2018, RefSeq2020, and the Kraken 2 Standard database are all whole-genome libraries, with no special modifications for use with 16S amplicon sequencing data.

Analysis pipeline and reference library pairings. We analyzed 136 mock community samples using a total of 11 distinct pairings of analysis tools and reference libraries: DADA2 only with SILVA, QIIME 2 with Greengenes and SILVA, Mothur only with SILVA (the default reference library), PathoScope using Greengenes, SILVA, RefSeq2018, and RefSeq2020, and Kraken 2 with its Standard library, SILVA, and Greengenes. While the SILVA database includes species-level taxonomic information for most of its representative 16S sequences, note that Mothur collapses feature counts into genus-level clades and thus does not make species-level calls. The adaptations of SILVA used for Kraken 2 and QIIME 2 did not provide species-level calls. Thus, only eight of the 11 pairings make species-level calls. The pairings and pipeline parameters are summarized in Table 1.

Tracking available taxonomic information for each OTU. A counts matrix was created from the results of each of the 11 pipeline/reference pairs for each operational taxonomic unit (OTU), denoised OTU, and feature. Each feature was assigned phylum, class, order, family, genus, species, and subspecies level information when available. For a given database, whenever a taxonomic label was missing, the lowest level taxonomy available for a feature was propagated using that database's taxonomic path, taking note of what granularity was

Analysis pipeline	Reference database(s) (lowest taxonomic level)	Pipeline parameters
DADA2	SILVA 138 (species)	Defaults; maxN = 0, maxEE = c(3, 3), truncQ = 2, rm.phix = TRUE, tlength = 0. For unpaired reads, maxEE = 3. For Kozich samples, tlength = c(240, 200). For Fouhy Ion Torrent samples, DADA2 function parameters set to HOMOPOLYMER_GAP_PENALTY = − 1 and BAND_SIZE = 32 with trimLeft = 15 for the filter/trim step
QIIME 2 using DADA2	Greengenes 13_8 (species), SILVA 138 (genus)	For most sequences, trimmed initial 6 bp. Kozich dataset, truncated to 240 bp for forward reads and 200 bp for reverse reads
Mothur	SILVA 138 (genus)	For pre.cluster() step, diffs = 2 for joined sequencing reads shorter than 250 bp; diffs = 3 for joined reads of length 250–349 bp; diffs = 4 for longer joined reads. For cluster.split(), taxlevel = 4, cutoff = 0.03
PathoScope 2.0	Greengenes 13_8 (species), SILVA 138 (species), RefSeq2018 (species), RefSeq2020 (species)	Bowtie2 alignment parameters: "–local -R 2 -N 0 -L 25 -i S,1,0.75 -k 10 –score-min L,100,1.28"
Kraken	Greengenes 13_8 (species), SILVA 138 (genus), Kraken Standard library (species)	Default parameters

Table 1. Summary table of all pipeline and reference database pairings, with information on non-default parameters chosen. There are 11 distinct pairings in all. For each pairing, a note was included as to the lowest taxonomic level call available. In sum, eight of 11 pairings were capable of making species-level calls. Any non-default parameters used for a given pipeline were also specified.

available (taxonomic best hit). For example, a feature assigned only as a member of the *Bacillales* order would be given the metadata: "phylum: *Firmicutes*, class: *Bacilli*, order: *Bacillales*, family: o_*Bacillales*, genus: o_*Bacillales*, species: o_*Bacillales*."

Assessing taxon sensitivity, read specificity, error, and diversity. Several metrics were used in assessing the overall quality and power of each 16S analysis pipeline and reference library at each taxonomic level. Results were independently evaluated at each taxonomic level. Any reads or features not assigned to a taxon at a given phylogenetic level were excluded from analysis, except where otherwise specified.

Taxon detection sensitivity. The metric of taxon detection sensitivity is defined here as the portion of expected taxa in a mock community sample detected by a given pipeline, at a minimum of 0.1% relative abundance. It essentially examines how often a given method can correctly determine an organism's presence in the mock community.

Read assignment specificity. Read assignment specificity is defined here as the portion of reads from a given sample assigned to taxa that are actually present in that sample's mock community. This is equivalent to 1 minus the portion of reads assigned to spurious taxa. This metric identifies the frequency of read assignment to incorrect organisms for a given method.

Error rate. The Normalized Root Mean Squared Error (NRMSE) was calculated as the root mean square error normalized with the assumption that the variance could be increasing given higher read counts. For each sample's results, given by the equation

$$\mathrm{NRMSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\frac{|w_i - t_i|^2}{\left(\frac{w_i+t_i}{2}\right)}}$$

where, for K taxa, w_i and t_i are respectively the measured and true read counts of taxon i. We evaluated the union of the expected and detected taxa for each sample, using $t_i = 0$ for theoretical taxa counts that were not actually measured in the mock community. All taxa that were absent from both the measured results and the true mock community were excluded (i.e., taxa which had relative abundance values of 0, both theoretical and measured).

Alpha diversity estimations. To assess each pipeline's ability to estimate the true alpha diversity within a sample regardless of accurate species identification, we calculated the log-fold change between the expected and the measured alpha diversity as measured by the Shannon index, the Simpson index, and the breakaway_nof1[5] index. The R package *vegan*[37] was used to calculate the Shannon and Simpson indices. The R package *breakaway* was used to calculate the breakaway_nof1[38] index, which predicts both the number of unobserved taxa and the number of true singletons based on the non-singleton frequency counts. Due to the alpha diversity metrics' sensitivity to library and count size differences[39], we converted the relative abundances of the mock community samples' ground truths to virtual sequencing libraries of 1,000,000 reads again using the *vegan*[37] package. A rarefaction depth of 10,000 reads per sample was used to normalize all samples and ground truth libraries.

Statistical methods for significance testing. A series of linear mixed-effects models (LMMs), coupled with post hoc least-square means tests and a Tukey multiple comparison correction, were used to determine which pipelines outperformed each other in sensitivity, specificity, error rates, and alpha diversity estimates.

LMMs were estimated using the lmer() function from the R package *lme4*[40], and post hoc comparisons were performed with the lsmeans() function from the R package *lsmeans*[41]. These LMMs examine the relevant performance metric as the measured variable, using the 136 mock community samples as a random effect and the pipeline/reference library pair as a fixed effect.

Results

Visual evaluation of species-level detection and abundance estimates. Figure 1 shows stacked bar charts of the results from the Kozich dataset for the ground truth versus all methods at the species level. Overall, pipelines using the Greengenes database (Kraken 2, QIIME 2, and Pathoscope) performed the worst in classifying species, followed by DADA2 paired with SILVA. PathoScope made the best use of the Greengenes database with the fewest misclassified reads and most correct species-level detection. Kraken 2 (paired with its Standard library) and PathoScope (paired with the RefSeq and SILVA libraries) performed best on these datasets. A more quantitative evaluation of these methods in the context of all samples follows.

Taxon detection sensitivity. At the genus level (Fig. 2A), DADA2 paired with SILVA was the least sensitive (mean = 0.67, SD = 0.35), followed collectively by methods which utilized the Greengenes (QIIME 2: mean = 0.73, SD = 0.16; Kraken 2: mean = 0.73, SD = 0.17; PathoScope: mean = 0.78, SD = 0.24; see Table S2 for p-values). When paired with the SILVA, RefSeq2018, or RefSeq2020 reference libraries, PathoScope was more sensitive at detecting genera than any other method, peaking when paired with the RefSeq 2018 reference library (mean = 0.88, SD = 0.14).

Generally, taxon detection sensitivity was lower at the species level than at the genus level (Fig. 2B). Methods using Greengenes had extremely low species-level sensitivities (QIIME 2: mean = 0.16, SD = 0.18; Kraken 2: mean = 0.19, SD = 0.13; PathoScope: mean = 0.28, SD = 0.21), as did DADA2 with SILVA (mean = 0.24, SD = 0.19). These were all significantly lower than all other methods (see Table S3 for pairwise p-values). Among those methods that used Greengenes, PathoScope was significantly more sensitive than either QIIME 2 ($p < 0.001$) or Kraken 2 ($p < 0.001$). The most sensitive method at the species level was PathoScope using the SILVA reference library (mean = 0.86, SD = 0.15), followed by PathoScope using RefSeq2018 (mean = 0.67, SD = 0.16). Only three species were not detected by PathoScope at a minimum of 0.1% relative abundance in any samples when using SILVA as a reference library; these were *Bifidobacterium adolescentis*, *Prosthecobacter fusiformis*, and *Clostridium beijerinckii*.

Read assignment specificity. At the genus level, the average read assignment specificity was generally lower for Kraken 2 with its Standard library (mean = 0.719, SD = 0.26); PathoScope and QIIME 2 with Greengenes (PathoScope: mean = 0.72, SD = 0.26; QIIME 2: mean = 0.73, SD = 0.28); and DADA2, Kraken 2, and Mothur with SILVA (DADA2: mean = 0.75, SD = 0.37; Kraken 2: mean = 0.75, SD = 0.2; Mothur: mean = 0.76, SD = 0.22).

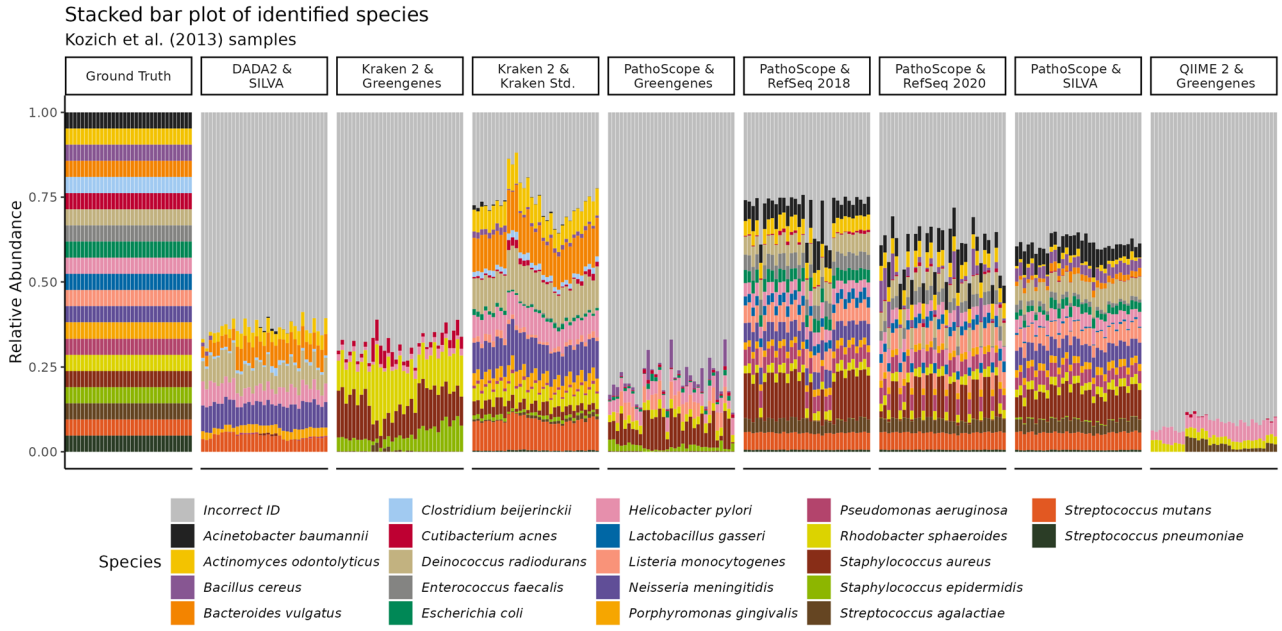

Figure 1. Expected versus measured relative abundances of mock bacteria. A stacked bar plot of the measured relative abundances of bacterial species in 33 samples from Kozich et al. These samples were all equimolar concentrations of 16S rDNA from 21 species, as shown in the 'Ground Truth' bar on the left. All reads assigned to bacterial species other than the 21 expected in the mock community are colored gray and are labeled "Incorrect ID". Mothur calls were not included as the pipeline does not make species-level calls, and the same is true of QIIME 2 and Kraken 2 paired with the SILVA database.

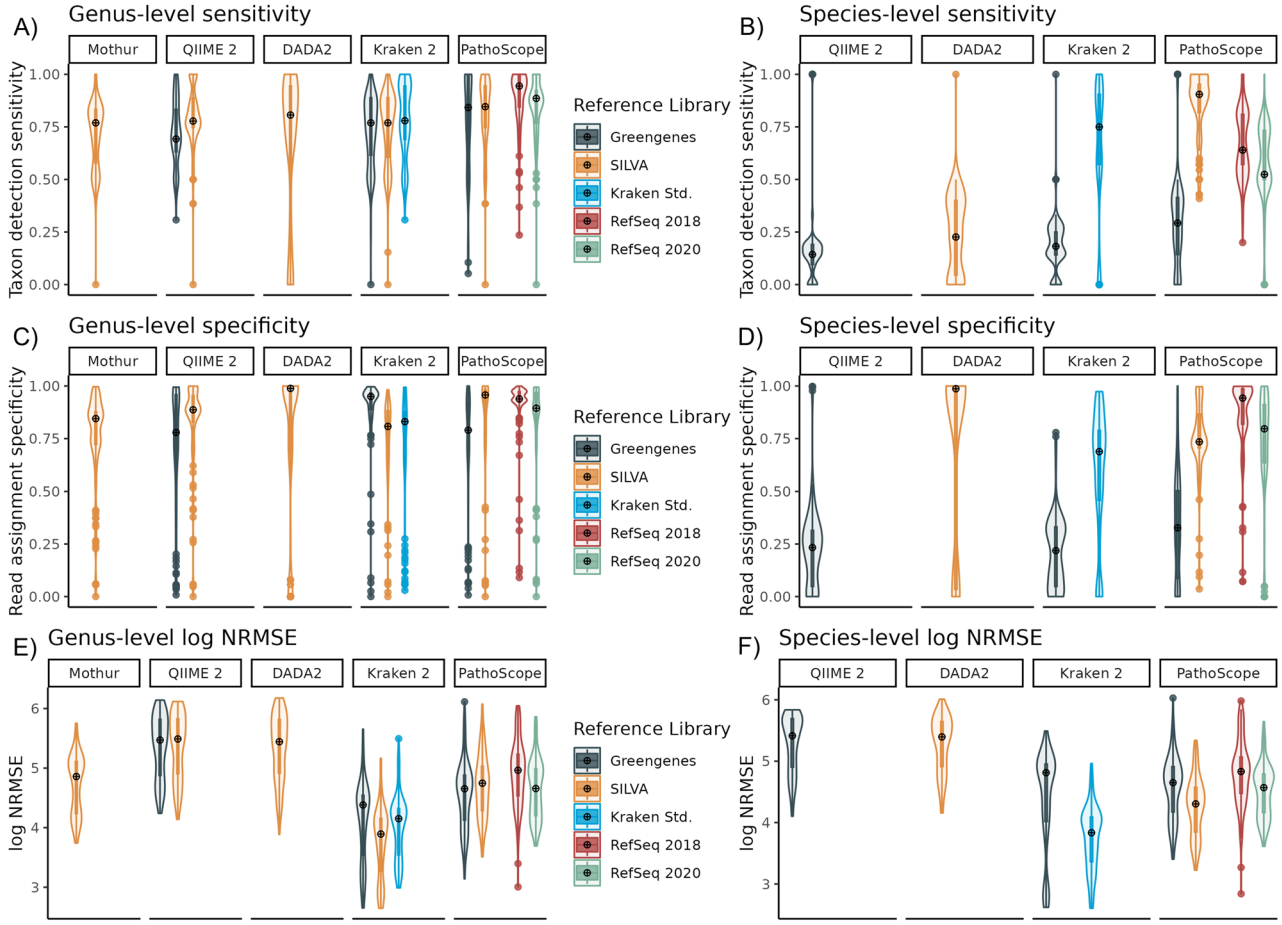

Figure 2. Taxon detection sensitivity of 16S analysis pipelines. Violin plots of the sensitivity, specificity, and log NRMSE of each analysis pipeline and reference library pair used to analyze 16S samples, calculated at the genus (**A**,**C**,**E**) and species (**B**,**D**,**F**) levels. Sensitivity is calculated as the portion of expected taxa in each mock community sample that was detected with least 0.1% relative abundance. Specificity is calculated as the portion of reads assigned to taxa that are expected to exist within each mock community.

(Fig. 2C). However, no overall trends arose in pairwise tests among pipelines and database pairings (see Table S4 for pairwise p-values). PathoScope with the RefSeq2018 library (mean = 0.91, SD = 0.15) and Kraken 2 with Greengenes (mean = 0.89, SD = 0.18) had the overall highest read assignment specificity.

At the species level, both Kraken 2 and QIIME 2 paired with Greengenes had the lowest read assignment specificity (Kraken 2: mean = 0.21, SD = 0.17; QIIME 2: mean = 0.23, SD = 0.2), which were significantly lower than all methods (see Table S5 for pairwise p-values). PathoScope, when paired with either the SILVA library (mean = 0.75, SD = 0.18), RefSeq2020 (mean = 0.75, SD = 0.24), or RefSeq2018 (mean = 0.86, SD = 0.18) was significantly more specific than QIIME 2 and Kraken 2 (Fig. 2D).

Normalized root mean-square error. Kraken 2 had the lowest error rates, measured as the log NRMSE of the raw reads, of all methods evaluated at the genus level regardless of the reference library used (SILVA: mean = 3.78, SD = 0.58; Standard: mean = 4.02, SD = 0.52, Greengenes: mean = 4.12, SD = 0.66). These were significantly lower than all other error rates (see Table S6 for pairwise comparison p-values). QIIME 2 had the highest genus-level NRMSE for both the SILVA and Greengenes libraries of all methods (SILVA: mean = 5.36, SD = 0.54; Greengenes: mean = 5.36, SD = 0.54), alongside DADA2 with SILVA (mean = 0.35, SD = 0.55; Fig. 2E).

At the species level, Kraken 2 also had the lowest log NRMSE for its Standard database and SILVA, which were better than all other methods (Standard: mean = 3.77; SD = 0.5; SILVA: mean = 3.95, SD = 0.55; see Table S7 for pairwise comparison p-values). This was followed by PathoScope for the SILVA database (mean = 4.28, SD = 0.48) and Kraken 2 using Greengenes (mean = 4.38, SD = 0.8). The worst NRMSE was again held by QIIME 2 with Greengenes and DADA2 using SILVA (QIIME 2: mean = 5.3, SD = 0.46; DADA2: mean = 5.29, SD = 0.45), which were significantly worse than all other methods (Fig. 2F).

Alpha diversity estimations. Out of all methods evaluated at the species level, Kraken 2 paired with Greengenes showed the greatest deviations from expected Shannon (deviation mean = 1.05, SD = 1.06) and Simpson (deviation mean = 0.25, SD = 0.27) alpha diversity indices, with significantly higher deviations than all other methods (Tukey-adjusted $p < 0.001$ in all pairwise comparisons). PathoScope generally matched the

true Shannon indices more closely than all other methods (RefSeq2020: deviation mean = 0.21, SD = 0.23; RefSeq2018: deviation mean = 0.27, SD = 0.28; Fig. 3A). The same trend held for the Simpson indices.

DADA2 reported the most closely matching log breakaway_nof1[38] indices, averaging significantly less deviation from the true number of species present than other methods (mean = 1.37, SD = 3.07; Tukey-adjusted $p < 0.001$ in all pairwise comparisons). On the other hand, Kraken 2 using its Standard library and SILVA frequently overestimated the number of species present by several orders of magnitude (Standard: mean = 6.17, SD = 1.82; SILVA: mean = 5.87, SD = 1.85), performing worse than all other methods (Fig. 3B).

Overall, no single pipeline or reference library performed the best in all evaluative metrics, but some holistic trends are present, especially at the species level. Figure 4B shows that sensitivity and specificity are correlated traits at the species level (Spearman's $r = 0.85$) and that PathoScope (regardless of reference library) and Kraken 2 (with its Standard library) dominate the upper right quadrant, where sensitivity and specificity are both high. In particular, PathoScope excels in both sensitivity and specificity when used with either SILVA or RefSeq2018. Similarly, Fig. 4C shows that error and alpha diversity estimated deviation are inversely correlated (Spearman's $r = -0.57$) and that no single method yields the lowest alpha diversity deviation and error rates. Trends are not well-defined at the genus level (Fig. 4A).

Discussion

Mock bacterial communities, either derived from spike-in DNA sequences or extracted from mixtures of bacterial cell monocultures, provide a semblance of a "ground truth" to assess 16S amplicon sequencing analysis methods. Ideally, knowing which species at what amounts should be present in any genuine microbiome sample would allow for accurate identification in every analysis. There are, of course, complications in sequencing experiments: technical bias and errors are introduced into samples at every step of the experiment until being safely sealed as bits in a FASTQ file on a server. Mock species' relative abundances may be affected by subtle variations in pipetting technique as spike-in DNA is aliquoted from individual sources. Spike-in DNA might be cloned from mutated DNA, or an early PCR error may have propagated through an entire commercial stock of nucleic acids. Different species of bacteria vary in lysing difficulty[42], causing some species to be underrepresented or even absent in the collected cDNA libraries from a plate[43]. While 16S amplification primers are designed to bind to universal conserved regions of the 16S rRNA gene, there is still clearly some amplification bias during PCR[44]. Contamination from reagents[45], local bacteria in the air, on gloves, or in a pipette tip box can further complicate matters. Thus, limitations of different experimental conditions and methods can dramatically affect the quality of results obtained from mock communities. Present sequencing errors and contamination suggest that amplicon reads will not identify with taxa as precisely as might a tidy, evenly distributed set of sequences drawn from a closed set of well-characterized species. It should then be evident that no analysis pipeline could theoretically exist to perfectly measure a mock community. Such a feat would require identifying only the expected species in their exact proportions, with no extraneous observations. As such, the proximate best method to analyze 16S amplicon sequencing data is one that identifies the makeup of the microbiome as truthfully as possible. Mock microbial communities can provide a level testing ground for existing tools to find their relative strengths and weaknesses in performance.

Of the pipelines tested, both QIIME 2 and Mothur were designed and built specifically for 16S amplicon sequencing analysis. Each has a suite of utility functions built to assist researchers in processing their data from the sequencer to differential abundance analysis and visualizations. Both are typically bundled with a dedicated bacterial 16S rRNA gene sequence database's reference library for alignment (i.e., Greengenes for QIIME 2, SILVA for Mothur). However, our results present strong evidence that PathoScope and Kraken 2 outperform QIIME 2, Mothur, and DADA2, even when comparing reads to identical reference databases. This phenomenon interestingly occurs despite Kraken 2 and PathoScope's status as more general whole genome sequencing and

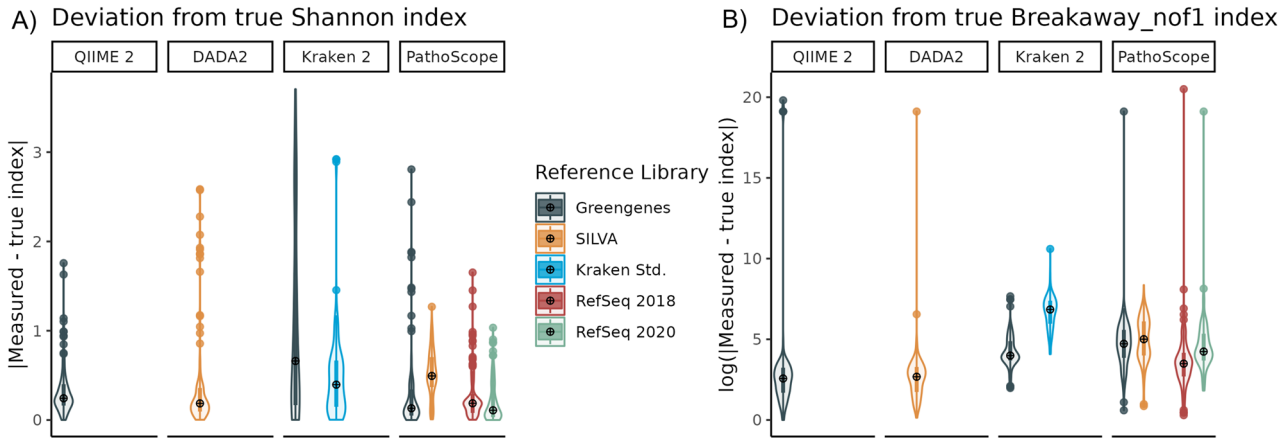

Figure 3. Deviation from true alpha diversity metrics at the species level. (**A**) The absolute difference between the measured Shannon alpha diversity index and the Shannon index value for the true mock community composition, and **B**) the log of the absolute difference between Breakaway_nof1 richness estimates and the true number of species present in each mock community. In both cases, values closer to 0 indicate more accurate estimation of the alpha diversity within a sample.

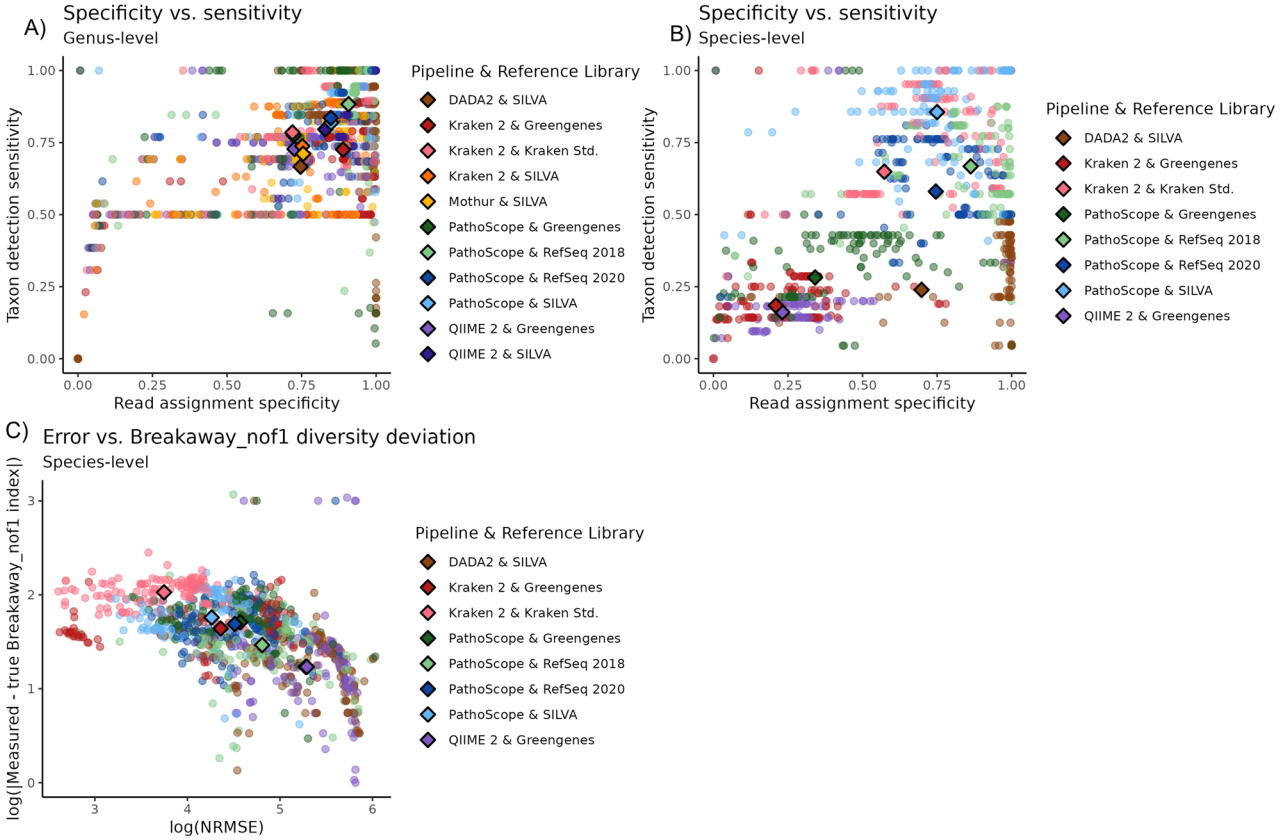

Figure 4. Combined quality of 16S analysis methods. Scatterplots of relative metrics for each 16S analysis pipeline, at the genus (**A**) and species (**B**,**C**) levels. Each point represents a single method's results when analyzing a single mock community sample. Points are colored by the analysis pipeline/reference library used. Centroids representing the mean values for each pipeline/reference library pair are marked with bolded diamonds.

metagenomics data tools. In pairwise comparisons, PathoScope is more sensitive and specific at taxon detection and has a lower error score than either DADA2, QIIME 2, or Mothur, and has comparable alpha diversity index estimates at both the genus and species levels. In general, SILVA's outperformance of Greengenes confirmed results found in previous benchmarks of 16S amplicon sequencing analysis methods[20–23,46]. This is likely due to several factors, including the small size of the Greengenes database in comparison to SILVA (99,000 versus 190,000), and the fact that this version of Greengenes had not been updated since 2013[47].

Kraken 2, when used with its Standard library, was rarely the top-performing analytic method in terms of sensitivity or specificity, although it was generally less error-prone than QIIME 2, Mothur, or any tool using Greengenes as a reference library. Kraken 2 has the added practical utility of being extremely fast and easy to use. Yet one constraint in analyzing Kraken 2 results is that they cannot be upsampled from a given taxon level, whereas PathoScope, QIIME 2, and Mothur all allow the tracing back of the taxonomic hierarchy of a given microbe. Both QIIME 2 and Mothur take advantage of naïve Bayes classifiers, which work most efficiently when trained on the specific region of the 16S rRNA gene amplified by PCR primers. Overall, PathoScope was the most sensitive in detecting taxa and specific in assigning reads, and the least error-prone tool when paired with either SILVA or RefSeq2018. However, it was not without limitations, as its computational expenses seemed to be an additional order of magnitude above those of other methods. This was evident from large interim SAM files (> 128 GB) and runtimes on the order of several hours, whereas Kraken 2 in particular took mere minutes. Issues aside, PathoScope is likely to outperform QIIME 2, DADA2, and Mothur in identification regardless of the database used. This finding partly results from PathoScope's Bayesian mixed modeling identification algorithm, which accounts for the possibility that multiple species can be present in the sample or that the target strain is not present in the reference database. PathoScope consistently outperformed Kraken 2 in most cases, although the difference was often slight and not statistically significantly better. Overall, these comparisons show that methods designed for general metagenomics analyses consistently outperform methods specifically designed for analyzing 16S data.

While many species are identifiable from their 16S rRNA gene sequence or a single hypervariable region, it is important to note that imperfect accuracy at this level is not solely a computational issue. For example, although the 16S rRNA gene is approximately 1550 bp long, the short sequencing reads obtained in most next-generation sequencing (NGS) only span about 250–500 bases and lack ideal resolution at the species level[48]. Compared to NGS, long-read sequencing technologies have been shown to perform better in classifying at the genus and species level[49,50]. We also observed differences in our study between the results of 15 samples sequenced with

Ion Torrent sequencing, as compared to samples from the same mock communities that were sequenced using Illumina Miseq. Furthermore, a major limitation to 16S amplicon studies is that some clades of bacteria exist with identical 16S DNA in the commonly sequenced V4 region. These clades of difficult-to-identify bacteria make up the bulk of Kraken 2's and PathoScope's incorrect calls. For example, *Bifidobacterium adolescentis* was nearly universally misclassified by all methods as other *Bifidobacterium* species, and *Prosthecobacter fusiformis* was frequently misidentified as *Prosthecobacter dejongeii*, a species with which it shares over 99% of its 16S DNA sequence[51]. Even further complications arise from many bacteria having several copies of the 16S rRNA gene, which may not be identical between operons within a genome[52]. This latter point may be in part why metagenomic methods such as Kraken 2 and PathoScope outperform specific methods such as QIIME 2 and Mothur, especially at the species level. The metagenomic methods are better designed to account for multiple 16S rRNA genes if present.

One of PathoScope's largest sources of error and lost taxon detection sensitivity calls when using the RefSeq2020 library comes from an apparent erroneous reference genome scaffold in the RefSeq representative genomes. In all mock community samples containing *Escherichia coli*, PathoScope with RefSeq2020 reported the presence of *Tumebacillus flagellates* at relative abundances tightly correlated with the expected values of *E. coli* (Pearson's $r = 0.959$). The circumstances strongly imply that reads actually originating from *E. coli* were incorrectly assigned to *T. flagellates*. *T. flagellates* is not even in the same phylum as *E. coli*, so the casual misassignment of reads between the species would be extremely unlikely based on 16S sequence similarity. Instead, a pairwise BLAST comparing *E. coli*'s 16S rRNA gene sequence to the *T. flagellates* scaffolds using the exact RefSeq entry that PathoScope had assigned those reads to (accession: NZ_JMIR01000093)[53] revealed that one *T. flagellates* scaffold had a 100% identity alignment over 911 bp. The finding possibly represents a case of horizontal gene transfer of the 16S rRNA gene, but it appears far more likely that *E. coli* contamination existed in the DNA library, which then was sequenced and assembled into *T. flagellates* scaffolds. On further study, it became apparent that this is merely one example of pervasive sequence contamination, meaning the accidental inclusion of sequences from other organisms or the misclassification of sequences, in public genome databases. This phenomenon has been recently explored in the NCBI RefSeq database[54–56]. The recent prevalence of high throughput and the accelerating low cost of next-generation sequencing (NGS) technologies has led to a rapid increase in published genomes available in the RefSeq libraries, although imperfect methods and protocols for sequencing data are contributing to high contamination rates. Human contamination in published genomes, while not a problem in 16S analyses, is a particularly frustrating problem when analyzing shotgun metagenomics data. Clearly, metagenomic read-mapping approaches such as Kraken 2 and PathoScope afford the potential for the development of novel quality control pipelines for RefSeq and other genome sequence databases.

The increasing prevalence of poor sequencing quality control helps to explain why the RefSeq 2018 libraries often performed better than the 2020 libraries. Many tools have been developed to identify and correct contamination errors in sequences and public databases[56–60], but this is an ongoing problem that demands additional filtering and correction efforts after directly retrieving libraries from the public repository. Given the higher specificity and sensitivity of PathoScope when using the 2018 RefSeq library over the 2020 library, we recommend using older RefSeq libraries until newer versions have been processed to remove contamination. Also of interest is the high accuracy of SILVA in its species calls when using PathoScope, even though it cannot be used to make such calls when used with QIIME 2, Mothur, or Kraken 2. SILVA also presents a viable alternative to the RefSeq libraries in avoiding contamination.

We note that in performing this benchmark, we sought to evaluate several common 16S amplicon sequencing analysis pipelines alongside metagenomics analysis pipelines. 16S pipelines were chosen based on performance as well as prevalence in previously published benchmarks. To identify metagenomics pipelines, we used a previously published metagenomics benchmarking paper[18]. Miossec et al. found that among the tested pipelines, PathoScope 2.0 and Kraken represented high sensitivity and specificity in the benchmark results. However, we emphasize that further comparison of other metagenomics analysis pipelines such as MetaMix[61], Centrifuge[62], and Metaxa2[63] should be conducted to analyze their differential performance, especially as new methods are developed and published.

Conclusion

DADA2, QIIME 2 and Mothur struggle to maintain accuracy at the genus level or granular species level in taxonomic analyses. Kraken 2, despite its primary purpose for whole genome sequencing metagenomics analyses, offers more power in analyzing 16S data without any increase in computational costs. PathoScope, while computationally more expensive, produces the most sensitive and accurate results of all evaluated pipelines when used on a diverse set of mock bacterial community samples. Analysis pipelines using SILVA as a reference library outperformed those using Greengenes significantly, and PathoScope using SILVA yielded the highest accuracies and sensitivities. While whole-genome reference libraries, such as Kraken 2's Standard or RefSeq's representative genomes, may provide some benefits over SILVA in terms of sensitivity, they may yield more spurious species-level calls. Based on the research conducted here with mock microbial communities, we recommend SILVA and RefSeq above other databases and strongly discourage usage of the Greengenes reference library for future analysis. While not included in our analysis due to release date, we do encourage users to try the Greengenes2[30] reference library as an phylogeny-based improvement over Greengenes. We also recommend PathoScope and Kraken 2 as fully capable, competitive options for conducting genus- and species-level 16S amplicon sequencing data analysis, in addition to outperforming other tools when using shotgun metagenomics data [18].

Data availability

Reference libraries used in analysis are available in the following GitHub repository: https://github.com/aubreyodom/16SBenchmarking.

Received: 14 September 2022; Accepted: 16 August 2023
Published online: 26 August 2023

References

1. Kumar, P. S. Microbiomics: Were we all wrong before?. *Periodontology 2000* **85**(1), 8–11 (2021).
2. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**(1), 5029. https://doi.org/10.1038/s41467-019-13036-1 (2019).
3. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119 (2017).
4. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**(7), 581–3. https://doi.org/10.1038/nmeth.3869 (2016).
5. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9 (2019).
6. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.* **7**(5), 335–336. https://doi.org/10.1038/nmeth.f.303 (2010).
7. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09 (2009).
8. Kopylova, E. *et al.* Open-source sequence clustering methods improve the state of the art. *mSystems* https://doi.org/10.1128/mSystems.00003-15 (2016).
9. Westcott, S. L. & Schloss, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**, e1487. https://doi.org/10.7717/peerj.1487 (2015).
10. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113 (2018).
11. Amir, A. *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* https://doi.org/10.1128/mSystems.00191-16 (2017).
12. Hong, C. *et al.* PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33. https://doi.org/10.1186/2049-2618-2-33 (2014).
13. Francis, O. E. *et al.* Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Res.* **23**(10), 1721–1729 (2013).
14. Byrd, A. L. *et al.* Clinical PathoScope: Rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinform.* **15**(1), 1–14 (2014).
15. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**(1), 257. https://doi.org/10.1186/s13059-019-1891-0 (2019).
16. He, Y. *et al.* Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. *Microbiome* **3**, 20. https://doi.org/10.1186/s40168-015-0081-x (2015).
17. Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364. https://doi.org/10.7717/peerj.5364 (2018).
18. Miossec, M. J. *et al.* Evaluation of computational methods for human microbiome analysis using simulated data. *PeerJ* **8**, e9688 (2020).
19. Miossec, M. J., Valenzuela, S. L., Mendez, K. N. & Castro-Nallar, E. Computational methods for human microbiome analysis. *Curr. Protoc. Microbiol.* **47**(1), 141–1417 (2017).
20. Dixit, K. *et al.* Benchmarking of 16S rRNA gene databases using known strain sequences. *Bioinformation* **17**(3), 377–391. https://doi.org/10.6026/97320630017377 (2021).
21. López-García, A. *et al.* Comparison of mothur and QIIME for the analysis of rumen microbiota composition based on 16S rRNA amplicon sequences. *Front. Microbiol.* **9**, 3010. https://doi.org/10.3389/fmicb.2018.03010 (2018).
22. Almeida, A., Mitchell, A. L., Tarkowska, A. & Finn, R. D. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* https://doi.org/10.1093/gigascience/giy054 (2018).
23. Lu, J. & Salzberg, S. L. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome.* **8**(1), 124. https://doi.org/10.1186/s40168-020-00900-2 (2020).
24. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05 (2006).
25. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-6. https://doi.org/10.1093/nar/gks1219 (2013).
26. Cole, J. R. *et al.* Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633-42. https://doi.org/10.1093/nar/gkt1244 (2014).
27. Lappan, R. *et al.* A microbiome case-control study of recurrent acute otitis media identified potentially protective bacterial genera. *BMC Microbiol.* **18**(1), 13. https://doi.org/10.1186/s12866-018-1154-3 (2018).
28. De Boeck, I. *et al.* Comparing the healthy nose and nasopharynx microbiota reveals continuity as well as niche-specificity. *Front. Microbiol.* **8**, 2372. https://doi.org/10.3389/fmicb.2017.02372 (2017).
29. Lapidot, R. *et al.* Nasopharyngeal dysbiosis precedes the development of lower respiratory tract infections in young Infants: A longitudinal infant cohort study. *medRxiv* **2**, 1 (2021).
30. McDonald, D. *et al.* Greengenes2 enables a shared data universe for microbiome studies. *bioRxiv* https://doi.org/10.1101/2022.12.19.520774 (2023).
31. Schoch, C. L. *et al.* NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* **01**(01), 2020. https://doi.org/10.1093/database/baaa062 (2020).
32. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189 (2016).
33. Lluch, J. *et al.* The characterization of novel tissue microbiota using an optimized 16S metagenomic sequencing pipeline. *PLoS ONE* **10**(11), e0142334. https://doi.org/10.1371/journal.pone.0142334 (2015).
34. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**(17), 5112–5120. https://doi.org/10.1128/AEM.01043-13 (2013).
35. Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. & Cotter, P. D. 16S rRNA gene sequencing of mock microbial populations: Impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* **16**(1), 123. https://doi.org/10.1186/s12866-016-0738-z (2016).

36. Karstens, L. *et al.* Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* https://doi.org/10.1128/mSystems.00290-19 (2019).
37. Oksanen, J. *et al. The Vegan Package: Community Ecology Package, Version 1.13-1*. https://www.veganr-forger-project.org (2008).
38. Willis, A. *Species richness estimation with high diversity but spurious singletons. arXiv preprint arXiv:160402598*. 2016;
39. Lundin, D. *et al.* Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity?. *Environ. Microbiol. Rep.* **4**(3), 367–372. https://doi.org/10.1111/j.1758-2229.2012.00345.x (2012).
40. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
41. Lenth, R. V. Least-squares means: The R package lsmeans. *J. Stat. Softw.* **69**, 1–33 (2016).
42. Gill, C., van de Wijgert, J. H., Blow, F. & Darby, A. C. Evaluation of Lysis methods for the extraction of bacterial DNA for analysis of the vaginal microbiota. *PLoS ONE* **11**(9), e0163148. https://doi.org/10.1371/journal.pone.0163148 (2016).
43. Boers, S. A., Jansen, R. & Hays, J. P. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur. J. Clin. Microbiol. Infect. Dis.* **38**(6), 1059–1070. https://doi.org/10.1007/s10096-019-03520-3 (2019).
44. Sze, M. A. & Schloss, P. D. The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *mSphere* https://doi.org/10.1128/mSphere.00163-19 (2019).
45. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**(1), 1–12 (2014).
46. Straub, D. *et al.* Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (Gene) amplicon sequencing pipeline. *Front. Microbiol.* **11**, 550420. https://doi.org/10.3389/fmicb.2020.550420 (2020).
47. Park, S.-C. & Won, S. Evaluation of 16S rRNA databases for taxonomic assignments using a mock community. *Genom. Inform.* **16**(4), e24 (2018).
48. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinform.* **17**(1), 1–8 (2016).
49. Nygaard, A. B., Tunsjø, H. S., Meisal, R. & Charnock, C. A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Sci. Rep.* **10**(1), 1–10 (2020).
50. Pearman, W. S., Freed, N. E. & Silander, O. K. Testing the advantages and disadvantages of short-and long-read eukaryotic metagenomics using simulated reads. *BMC Bioinform.* **21**(1), 1–15 (2020).
51. Lee, J., Park, B., Woo, S. G. & Park, J. *Prosthecobacter algae* sp. nov., isolated from activated sludge using algal metabolites. *Int. J. Syst. Evol. Microbiol.* **64**(Pt 2), 663–667. https://doi.org/10.1099/ijs.0.052787-0 (2014).
52. Louca, S., Doebeli, M. & Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**(1), 41. https://doi.org/10.1186/s40168-018-0420-9 (2018).
53. Wang, Q. *et al. Tumebacillus flagellatus* sp. Nov., an α-amylase/pullulanase-producing bacterium isolated from cassava wastewater. *Int. J. Syst. Evol. Microbiol.* **63**(Pt 9), 3138–3142. https://doi.org/10.1099/ijs.0.045351-0 (2013).
54. Lupo, V. *et al.* Contamination in reference sequence databases: Time for divide-and-rule tactics. *Front. Microbiol.* **12**, 755101. https://doi.org/10.3389/fmicb.2021.755101 (2021).
55. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**(6), 954–960. https://doi.org/10.1101/gr.245373.118 (2019).
56. Steinegger, M. & Salzberg, S. L. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**(1), 115. https://doi.org/10.1186/s13059-020-02023-1 (2020).
57. Lu, J. & Salzberg, S. L. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* **14**(6), e1006277. https://doi.org/10.1371/journal.pcbi.1006277 (2018).
58. Cornet, L. & Baurain, D. Contamination detection in genomic data: More is not enough. *Genome Biol.* **23**(1), 60. https://doi.org/10.1186/s13059-022-02619-9 (2022).
59. De Simone, G. *et al.* Contaminations in (meta)genome data: An open issue for the scientific community. *IUBMB Life* **72**(4), 698–705. https://doi.org/10.1002/iub.2216 (2020).
60. Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* **19**(1), 1–10 (2018).
61. Morfopoulou, S. & Plagnol, V. Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics* **31**(18), 2930–2938 (2015).
62. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**(12), 1721–1729 (2016).
63. Bengtsson-Palme, J. *et al.* METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* **15**(6), 1403–1414 (2015).

Author contributions

T.F., E.C., K.A.C. and W.E.J. conceived the study design. T.F. and A.R.O. conducted the research study, performed all computational work, wrote the main manuscript text, and prepared figures and tables. W.E.J. also wrote the main manuscript text. All authors read and approved the final manuscript.

Funding

T.F. and W.E.J. were supported in part by the NIH under grant R01GM127430. A.R.O. and W.E.J. were supported in part by the NIH under grant R21AI154387.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-40799-x.

Correspondence and requests for materials should be addressed to W.E.J.