Microbiome

METHODOLOGY Open Access

# Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets

Isabel F. Escapa[1,2,3†], Yanmei Huang[1,2†], Tsute Chen[1,2], Maoxuan Lin[1], Alexis Kokaras[1], Floyd E. Dewhirst[1,2] and Katherine P. Lemon[1,3,4,5*]

## Abstract

**Background:** The low cost of 16S rRNA gene sequencing facilitates population-scale molecular epidemiological studies. Existing computational algorithms can resolve 16S rRNA gene sequences into high-resolution amplicon sequence variants (ASVs), which represent consistent labels comparable across studies. Assigning these ASVs to species-level taxonomy strengthens the ecological and/or clinical relevance of 16S rRNA gene-based microbiota studies and further facilitates data comparison across studies.

**Results:** To achieve this, we developed a broadly applicable method for constructing high-resolution training sets based on the phylogenic relationships among microbes found in a habitat of interest. When used with the naïve Bayesian Ribosomal Database Project (RDP) Classifier, this training set achieved species/supraspecies-level taxonomic assignment of 16S rRNA gene-derived ASVs. The key steps for generating such a training set are (1) constructing an accurate and comprehensive phylogenetic-based, habitat-specific database; (2) compiling multiple 16S rRNA gene sequences to represent the natural sequence variability of each taxon in the database; (3) trimming the training set to match the sequenced regions, if necessary; and (4) placing species sharing closely related sequences into a training-set-specific supraspecies taxonomic level to preserve subgenus-level resolution. As proof of principle, we developed a V1–V3 region training set for the bacterial microbiota of the human aerodigestive tract using the full-length 16S rRNA gene reference sequences compiled in our expanded Human Oral Microbiome Database (eHOMD). We also overcame technical limitations to successfully use Illumina sequences for the 16S rRNA gene V1–V3 region, the most informative segment for classifying bacteria native to the human aerodigestive tract. Finally, we generated a full-length eHOMD 16S rRNA gene training set, which we used in conjunction with an independent PacBio single molecule, real-time (SMRT)-sequenced sinonasal dataset to validate the representation of species in our training set. This also established the effectiveness of a full-length training set for assigning taxonomy of long-read 16S rRNA gene datasets.

(Continued on next page)

* Correspondence: Katherine.lemon@bcm.edu
†Isabel F. Escapa and Yanmei Huang contributed equally to this work.
[1]Forsyth Institute (Microbiology), Cambridge, MA, USA
[3]Department of Molecular Virology & Microbiology, Alkek Center for Metagenomics & Microbiome Research, Baylor College of Medicine, Houston, TX, USA
Full list of author information is available at the end of the article

(Continued from previous page)
**Conclusion:** Here, we present a systematic approach for constructing a phylogeny-based, high-resolution, habitat-specific training set that permits species/supraspecies-level taxonomic assignment to short- and long-read 16S rRNA gene-derived ASVs. This advancement enhances the ecological and/or clinical relevance of 16S rRNA gene-based microbiota studies.

**Keywords:** Training set, Naïve Bayesian RDP Classifier, Species-level taxonomy, 16S rRNA gene, V1–V3, Habitat-specific database, Microbiome, eHOMD, Nasal, Aerodigestive tract

## Background

In microbiota studies of most ecosystems and/or habitats, achieving species- or strain-level identification of constituents improves the ecological and/or clinical relevance of the results compared with genus-level identification. For example, species-level identification is often critically important for host-associated microbial communities because these communities frequently include commensal and pathogenic species of the same genus, e.g., [1, 2], with the caveat that strain-level variation in pathogenicity also exists, e.g., [3]. Additionally, some microbial genera include species that are site specialists and inhabit distinct niches of a given environment [4]. High-throughput close-to-full-length 16S rRNA gene sequencing (e.g., circular consensus sequences from PacBio single molecule, real-time (SMRT) sequencing) and metagenomic whole genome sequencing (WGS) hold promise for species- and strain-level microbiota studies. However, the easier accessibility, and lower cost, of 16S rRNA gene short-read sequencing makes population-scale (i.e., thousands of samples) molecular epidemiological studies of bacterial microbiota of humans, other animals, plants, and the environment broadly feasible now. A caveat to this is that the majority of published short-read 16S rRNA gene sequencing studies use read clustering at a percent similarity that constrains resolution to the genus level, i.e., 97% identity. Indeed, recent reviews on best practices and benchmarking for 16S rRNA gene microbiota studies focus on genus-level operational taxonomic unit (OTU) analysis, e.g., [5–7]. However, newer algorithms, many of which are not based on similarity thresholds, allow single-nucleotide resolution and can resolve 16S rRNA gene short-read sequences into species- or strain-level phylotypes, usually called amplicon sequence variants (ASVs) (e.g., MED (minimal entropy decomposition) [8, 9], DADA2 (divisive amplicon denoising algorithm) [10, 11], and UNOISE2 [12, 13], among others [14, 15]). Another limitation of short-read 16S rRNA microbiome studies is that the choice of the 16S rRNA gene region(s) sequenced places an upper bound on the degree of species-level resolution that is achievable within a dataset [16–19]. Therefore, it is critical to determine which regions provide the most information for distinguishing bacterial species that are common to that specific ecosystem. For the habitats within the human aerodigestive tract, i.e., the nasal passages, sinuses, throat, mouth, esophagus, as well as the lower respiratory tract, we previously showed that many more taxa are distinguishable at the species level with the V1–V3 region than with the commonly used V3–V4 region of the 16S rRNA gene [20]. Therefore, we have developed a method that, by combining the "reusability, reproducibility, and comprehensiveness" of ASVs, per Callahan and colleagues [11, 13], and the selection of highly informative regions of the 16S rRNA gene, maximizes 16S rRNA gene short-read sequencing potential to achieve sub-genus level resolution taxonomic assignment.

Microbial databases encompassing broad phylogenetic diversity, such as SILVA [21, 22], RDP [23], and Greengenes [24], serve the key role of being applicable to myriad different habitats. However, this valuable breadth comes with the trade-off of inclusion of taxonomically misannotated 16S rRNA gene sequences. For example, Edgar estimated annotation error rates as high as ~ 10–17% in these comprehensive databases [25]. SILVA and RDP continue to undergo regular updates and contain a broadly expansive and comprehensive record of 16S rRNA gene sequences from all explored habitats, whereas Greengenes was last updated in 2013. For habitats that have yet to be deeply interrogated, the access to this breadth outweighs the risk of misclassification due to annotation error. However, once a habitat is sufficiently explored through sequencing, creation of a habitat-specific database enables accurate fine-level phylogenetic resolution for taxonomic assignment to ASVs [20, 26–35]. Existing habitat-specific databases are constructed with different methods and can be used to assign taxonomy via different approaches, as in the following three examples. First, there are stand-alone habitat-specific databases consisting of curated collections of close-to-full-length 16S rRNA gene sequences compiled both from other repositories and by generating new sequences from the habitat of interest, e.g., eHOMD for the human aerodigestive tract [20, 26, 36], HITdb for the human gut [29], and RIM-DP for rumen [28]. Second, custom addition of compiled sequences from a specific habitat of interest can be used to augment a broad

general database, e.g., HBDB for honey bee [27], DictDB for termite and cockroach gut [33], SILVA19Rum for rumen [35], and MiDAS for activated sludge [30, 32]. Third, both a general and a habitat-specific database can be combined in the same pipeline, e.g., for human-associated genera with pathogenic members [37], for freshwater (FreshTrain with the TaxAss workflow [34]) and the multistage blastn workflow with eHOMD [38]. Many of these databases are formatted for use as training sets to train classifier algorithms for taxonomy assignment.

The naïve Bayesian RDP Classifier [39] is one of several effective algorithms for assigning taxonomy. (For benchmarking and comparison with other methods, see [40–43].) This type of supervised learning algorithm requires a training set, which is a set of input-output examples to learn a function that can be used to make predictions [44]. In this case, sequences are input and taxonomic assignments are output. Properly formatted versions of the broad 16S rRNA gene databases SILVA, RDP, and Greengenes are available to train the most popular implementations of the naïve Bayesian RDP Classifier. The quality of the training set strongly influences taxonomic assignment and habitat-specific training sets have been developed to increase accuracy of taxonomic assignments [27, 33, 40–42, 45]. However, the resolution of available training sets is mostly limited to the genus level. An exception is the manually curated subset of the Greengenes database corresponding to 89 clinically relevant bacterial genera that was used to assign species-level taxonomy of full-length 16S rRNA gene sequences of clinical isolates [46]. Notwithstanding, species-level taxonomy assignment of short-read 16S rRNA gene datasets remains a challenge.

We hypothesized that we could develop a method to rapidly generate a habitat-specific training set to leverage the strength of the naïve Bayesian RDP Classifier to consistently achieve species- or supraspecies (i.e., subgenus)-level taxonomic assignment of ASVs starting from a locally comprehensive, phylogeny-based, high-resolution (i.e., separated at ≥ 98.5% identity) set of curated reference sequences with distinct taxonomic names. Here, we show that the use of the naïve Bayesian RDP Classifier with a training set in which each taxon is represented by a collection of highly similar sequences that captures the natural variability of each species resulted in accurate species-level taxonomic assignment of short-read and long-read sequences of the 16S rRNA gene. This represents a methodological advancement. Our systematic approach for generating training sets is applicable to any ecosystem/habitat of interest and is summarized in Fig. 1. This approach requires compiling high-quality close-to-full-length 16S rRNA gene datasets from a habitat (Fig. 1a). These compiled datasets are then used to identify curated reference sequences to build a 16S rRNA gene database (Fig. 1b) from which training sets for that habitat are derived (Fig. 1c).

To test our hypothesis, we developed and validated short- and long-read training sets for the microbiota of the human aerodigestive tract (mouth, nasal passages, sinuses, throat, and esophagus) using our expanded Human Oral Microbiome Database (eHOMD). This database was originally created and later expanded to serve as a resource for the community of investigators generating datasets to study habitats within the human aerodigestive tract [20, 26, 36]. In addition to 16S rRNA gene reference sequences (eHOMDrefs), it also includes genomic and proteomic data. (It also works well for the lower respiratory tract [20].) The lack of proper taxonomical representation in traditional databases is a challenge in predicting taxonomic assignments [43]. A strength of the eHOMD is that, by placing 16S rRNA gene reference sequences for each human microbial taxon (HMT) on a phylogenetic tree (http://ehomd.org/index.php?name=HOMD&show_tree=_), as-yet unnamed or uncultivated species are defined based on sequence identity and added to the phylogeny using a provisional naming scheme that permits taxonomic assignment for cross-study comparison [26]. Also, sequences that are misnamed in other databases are easily identified and given a correct designation in eHOMD. Furthermore, each HMT in eHOMD is represented by one to six highly curated eHOMDrefs to account for intraspecies variability across different strains and dissimilar 16S rRNA genomic copies within individual strains [20]. Another key strength of eHOMD is that it is locally comprehensive often allowing approximately 95% of sequences from V1 to V3 aerodigestive tract datasets to be assigned to the taxonomy [20].

## Results

### Compiling closely related sequences for each taxon in a training set improves the accuracy of species-level taxonomic classification

Genus-level taxonomic assignment is not an inherent limitation of the naïve Bayesian RDP Classifier. Rather, taxonomic assignment to 16S rRNA gene short reads is limited by both the resolution to which sequences in datasets are distinguished and by the nature of the training set used. The former is overcome by using approaches such as oligotyping/MED [8, 9], DADA2 [10, 11], or UNOISE2 [12, 13] to resolve sequence variants at single-nucleotide scale. We hypothesized that the limitations inherent in training sets could also be overcome.

The naïve Bayesian RDP Classifier algorithm indicates that a training set with a larger number of sequences representing each taxon will result in more confident taxonomic assignment [39]. Based on the conditional

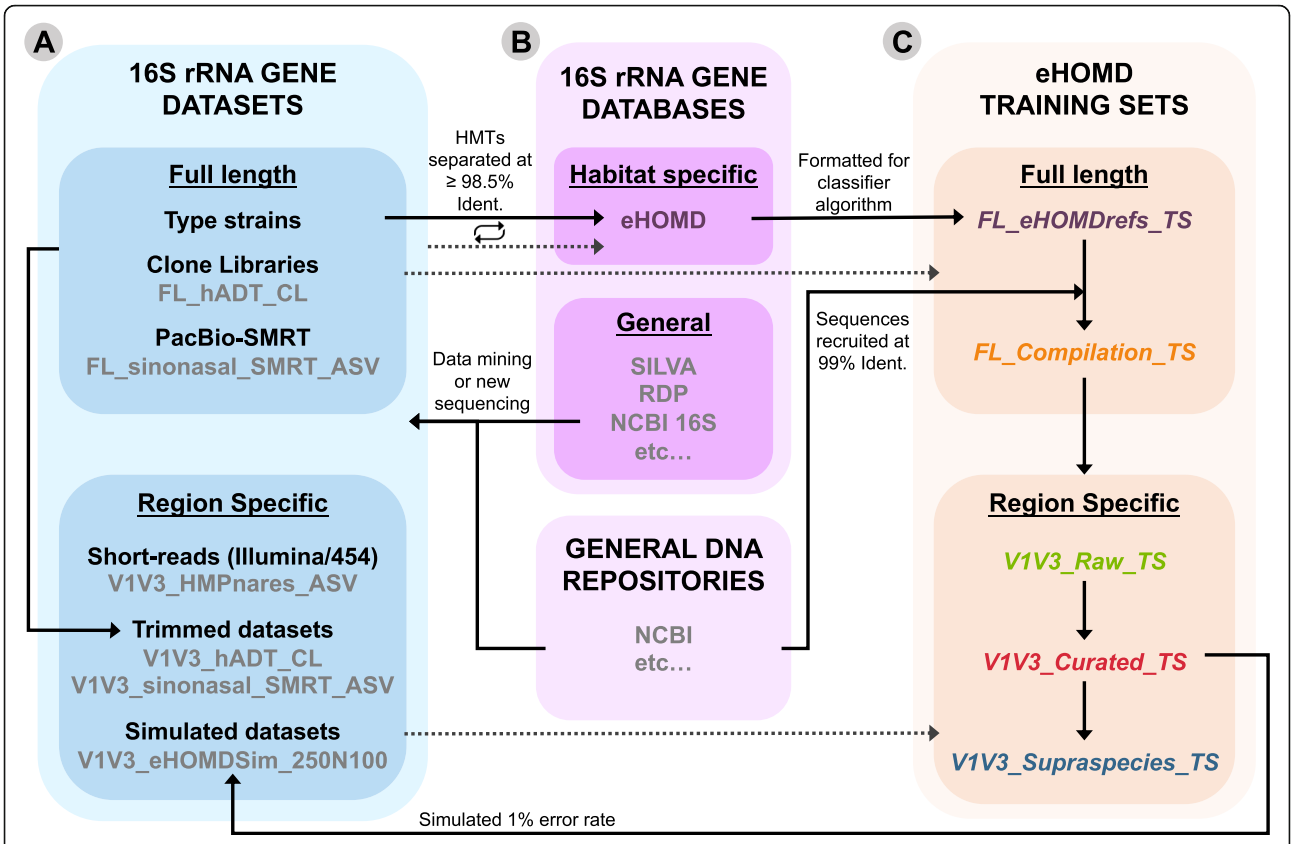


**Fig. 1** Relationships between the datasets, databases, and training sets in constructing training sets for a specific habitat: the human aerodigestive tract. **a** Datasets gathered from public repositories or obtained by sequencing of new samples are used to explore the 16S rRNA gene diversity of the habitat of interest. These include both 16S rRNA full-length sequences and region-specific short-read sequences used for method validation or benchmarking. **b** A curated habitat-specific full-length 16S rRNA gene reference database is assembled and expanded in an iterative way by selecting from those datasets representative sequences for both named and as-yet unnamed or uncultivated species (i.e., HMTs in eHOMD), and placing them in a phylogenetic tree (See Figure 1 in [20]). **c** Training sets are derived from the taxonomical hierarchy of the habitat-specific database and enhanced by the following steps: compiling multiple 16S rRNA gene sequences to represent the natural sequence variability of each taxon, trimming the training set to match the sequenced region/s, and placing species sharing closely related sequences into a supraspecies taxonomic level. Datasets in gray are the specific examples used for the construction of the eHOMD derived training sets described here. Solid arrows indicate where the sequences described come from and dotted arrows indicate when datasets were used for validation or benchmarking

probability for a member of a taxon ($T$), the higher the occurrence of a given distinguishing "k-mer" (word or $w_i$) in the training set, the greater the confidence with which assignment of that taxon is made, i.e., more sequences can be classified unambiguously. Thus, as the number of sequences ($M$) for each taxon in the training set increases (Fig. 2a vs. b), the number of accurate assignments should increase. (See Additional file 1 for a more detailed explanation.) In support of this, Werner and colleagues found more sequence counts result in more assignments [45]. Therefore, to systematically increase $M$, we used each of the reference sequences in eHOMD, i.e., the eHOMDrefs, as bait to capture closely matching, publicly available sequences and combined the resulting compilation of sequences for each taxon into a close-to-full-length compilation training set (FL_Compilation_TS) that reflects the currently known 16S rRNA gene sequence variability for each taxon, both natural and sequencing-error derived (Additional file 2).

To evaluate the performance of the FL_Compilation_TS and for further method optimization, we created the simulated dataset V1V3_eHOMDSim_250N100 by introducing a 1% error rate in a V1–V3 trimmed version of our training set sequences. (See "Methods" section for details and Additional file 3 for the dataset itself.) This simulated dataset aims to mimic real sequencing short-reads similar to the ones that can be obtained with our method for achieving highly informative 16S rRNA gene V1–V3 region sequencing data using Illumina MiSeq, which is detailed in Additional file 4. We then assessed the percentage of reads in the simulated dataset V1V3_eHOMDSim_250N100 that classified at the species-level

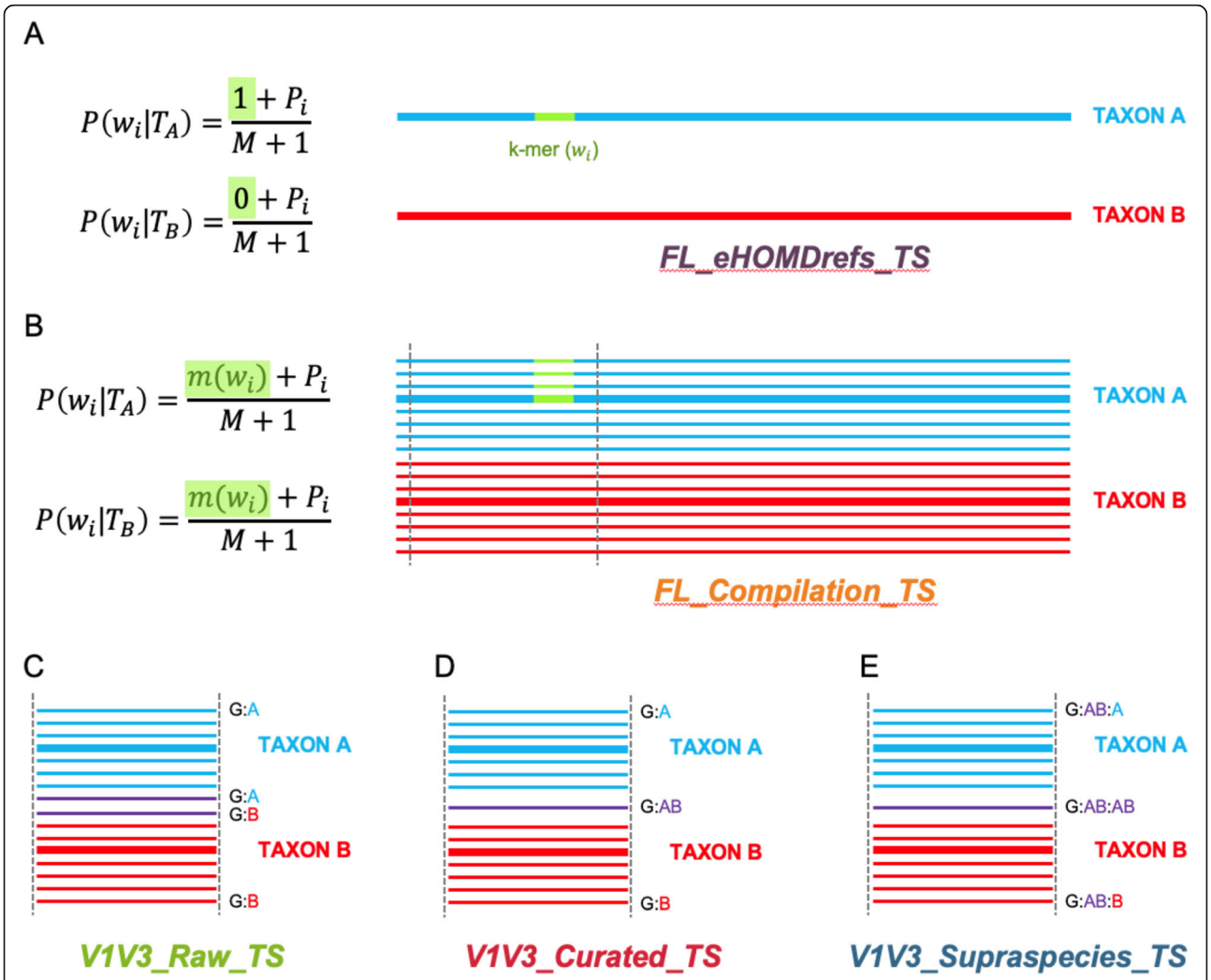


**Fig. 2** Schematic representation of the steps to generate sequential habitat-specific training sets. **a** The FL_eHOMDrefs_TS training set contains all full-length eHOMDrefs (thick lines) from eHOMDv15.1 together with their respective taxonomic assignment. When only one read represents each taxon ($M = 1$), a given distinguishing k-mer (green fragment) can only be either present (1) or absent (0). **b** A higher number of sequences per taxon ($M$) allows for better resolution on the assignment, with the presence of a given distinguishing k-mer ($w_i$) across each cluster of reads (green fragments) being represented as a proportion ($m$) out of the total number of reads in that taxon ($M$). Therefore, to better represent the known sequence diversity of the 16S rRNA gene(s) for each taxon, the training set FL_Compilation_TS includes clusters of sequences (thin lines) recovered from the NCBI nonredundant nucleotide (nr/nt) database that matched with 99% identity and ≥ 98% coverage (see methods) to each eHOMDref (thick line). **c** The training set V1V3_Raw_TS is a V1–V3 trimmed version of the FL_Compilation_TS training set. The schematic illustrates how trimming to this region leads to identical reads (purple lines) having two different taxonomic designations. Here, G is genus and species are labeled as A or B. **d** To construct the V1V3_Curated_TS training set, identical V1–V3 sequences in the V1V3_Raw_TS training set were collapsed into one. If identical sequences came from more than one taxon (purple), species-level names of all taxa involved were concatenated (AB). **e** The V1V3_Supraspecies_TS training set includes the same sequences that the V1V3_Raw_TS training set; however, the headers in the fasta file include the supraspecies taxon (AB) as an extra level between the genus (G) and species taxonomic levels (A, B, or AB), as illustrated here

at incremental bootstrap values from 50 to 100 using the training set FL_Compilation_TS (Fig. 3a, orange bars) compared with the training set FL_eHOMDrefs_TS (Fig. 3a, purple bars), which consists of only the eHOMDrefs. Doing this, we observed an increase in the percentage of reads that classified at species-level with the compilation TS except at a bootstrap value of 100. We postulate that the additional sequences classified with the training set FL_eHOMDrefs_TS at a bootstrap of 100 are misclassified sequences. A higher rate of misclassification is expected with the training set FL_eHOMDrefs_TS, since it includes only a few representative sequences for each taxon. Therefore, we next assayed for any misclassifications. At each bootstrap threshold, the percentage of reads that were misclassified (i.e., reads for which the assigned taxonomic identity was different than the known identity of the original sequence from which the simulated read was derived) was at least 50% lower using the training set FL_Compilation_TS (Fig. 3b,

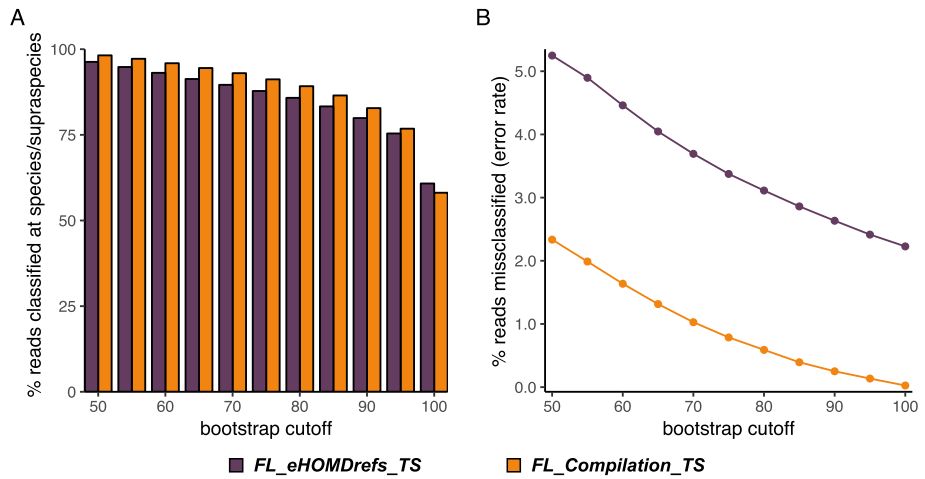


**Fig. 3** The FL_Compilation_TS training set provides higher classification percentages with a lower error rate. **a** The percentage of eHOMD-derived simulated reads classified using the FL_eHOMDrefs_TS training set (purple) versus the FL_Compilation_TS training set (orange). **b** The percentage of classified reads that were misclassified (i.e., reads for which the assigned taxonomic identity was different than the known identity of the original sequence from which the simulated read was derived). The naïve Bayesian RDP Classifier was used with bootstrap values ranging from 50 to 100

orange line) than with the training set FL_eHOMDrefs_TS (Fig. 3b, purple line). Thus, classification of the dataset V1V3_eHOMDSim_250N100 showed a reduced error rate and increased confidence level when using a training set consisting of a compilation of closely related sequences rather than one consisting of only one or a few reference sequences for each taxon.

## Moving toward an appropriate short-read fragment training set

The training set FL_Compilation_TS consisted of close-to-full-length 16S rRNA gene sequences. We hypothesized that the presence of k-mers outside the V1–V3 region in the training set might lead to misclassifications when using it with a V1–V3 region dataset. Supporting this, when using a training set based on a large general database, trimming reference sequences to match the sequenced region increases the number of sequences that are assigned taxonomy [45]. Therefore, we trimmed the sequences in the training set FL_Compilation_TS to cover only the V1–V3 region generating training set V1V3_Raw_TS (Fig. 2c). Using this, there was no gain in the percentage of reads in the simulated dataset V1V3_eHOMDSim_250N100 classified to the species-level (Fig. 4a, green bar). Moreover, we observed an increase in the percentage of misclassified reads (Fig. 4b, green line), i.e., the assignment accuracy decreased. Therefore, we next determined why the use of appropriate short-read fragments in this training set paradoxically decreased confidence and accuracy of species-level assignment and resolved the issue.

## Combining closely related, indistinguishable taxa into supraspecies decreases the error rate for a short-read training set

Considering the possible explanations for the above paradox, we realized that taxa with distinct full-length 16S rRNA gene sequences can have identical V1–V3 sequences. In silico, using only V1–V3, 37 of the ~770 species-level taxa in eHOMD are no longer distinguishable from at least one other species at 100% identity [20]. Therefore, we hypothesized that the observed loss in accuracy using training set V1V3_Raw_TS was due to identical sequences with more than one species name, e.g., *Veillonella parvula* and *Veillonella dispar* [20]. To solve this problem, we removed duplicate sequences and assigned a combined name, i.e., a training-set-specific supraspecies name such as *Veillonella* parvula_dispar, to the remaining unique sequence (Fig. 2d). We note that the term supraspecies is not intended to be a valid taxonomic label, since it is dependent on the database and can vary for different short-read 16S rRNA gene regions, i.e., is training-set specific. This resulted in the training set V1V3_Curated_TS, which showed improved accuracy compared with both the FL_Compilation_TS and the V1V3_Raw_TS (Fig. 4b, red line). However, this improvement came at a cost of a 0.7 to 4.4 % decrease in the reads assigned supraspecies- or species-level taxonomy at each bootstrap threshold (Fig. 4a, red bar). This trade-off can be illustrated by graphing at each bootstrap threshold the percentage of reads from the simulated V1V3_eHOMDSim_250N100 dataset that were misclassified versus confidently assigned species-level taxonomy using the V1V3_Curated_TS with the naïve Bayesian RDP Classifier (Fig. 4c).

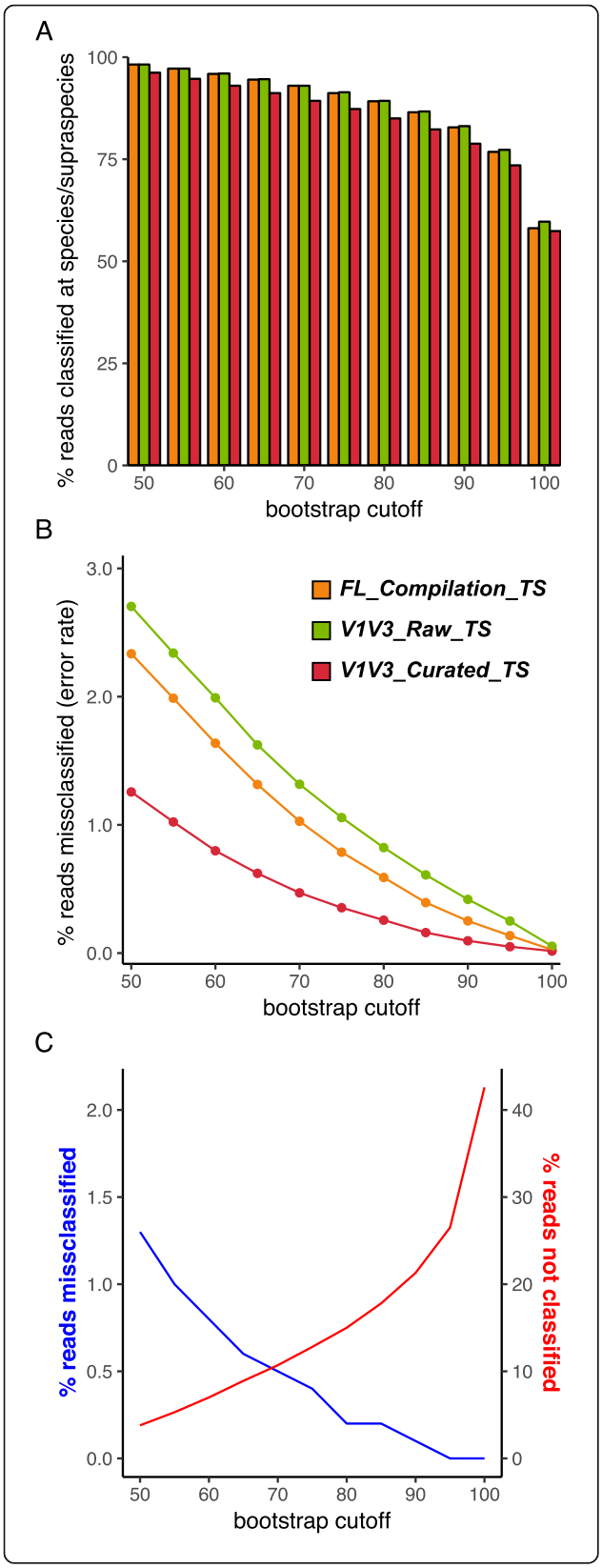


**Fig. 4** Trimming the training set to the specific sequenced region further reduces the error rate. **a** The percentage of eHOMD-derived simulated reads classified at species level using the FL_Compilation_TS (orange) training set compared to subsequent trimmed versions V1V3_Raw_TS (green) and V1V3_Curated_TS (red). **b** The percentage of classified reads that were misclassified with each of these three training sets. **c** This graph, which is specific to the eHOMD training set construction (V1V3_eHOMDSim_250N100 dataset), indicates how researchers can determine the bootstrap value to use with the naïve Bayesian RDP Classifier by deciding an acceptable level of the % of reads misclassified (blue line; e.g., 0.5%) and/or of the % of reads that are not classified (red line). The naïve Bayesian RDP Classifier was used with bootstrap values ranging from 50 to 100

### Maximizing performance of the naïve Bayesian RDP Classifier for subgenus-level taxonomic assignment of short-read sequences requires both insertion of supraspecies as a taxonomic level and setting a threshold bootstrap

Two steps were required to reap the benefits of using the supraspecies definition. First, formally inserting supraspecies as a training-set-specific taxonomic level between genus and species in the name header for each sequence in the training set, which yielded the training set V1V3_Supraspecies_TS (Fig. 2e; Additional file 5). Second, establishing the bootstrap cut-off at which a sequence is not assigned at the species level so that the naïve Bayesian RDP Classifier will then default to the supraspecies level, rather than defaulting to genus, allowing for a higher percentage of reads to be assigned (Fig. 5a, blue bar). This latter step preserves subgenus-level information encoded in the ASVs. This choice inevitably involves a trade-off between accuracy and percentage of reads classified below genus level, e.g., see Fig. 4c. For our purposes, we chose a conservative bootstrap value of 70 (Fig. 5b, blue line). With the simulated data, for which the truth is known, this gave an error rate of around 0.05%. Of note, although one common bootstrap setting for the naïve Bayesian RDP Classifier is 50, we use a more conservative value for species-level taxonomic assignment with the V1V3_Supraspecies_TS.

### Using an independent sinonasal dataset validated the performance of the eHOMD training set

We next validated the performance of the eHOMD training set against a dataset of close-to-full-length 16S rRNA gene sequences from sinonasal samples. Very importantly, these sequences are independent of the sequences recruited to build the training set. Earl and colleagues performed PacBio SMRT sequencing of close-to-full-length 16S rRNA genes from multiple sinonasal samples from each of 12 adults without sinus inflammation who were undergoing removal of a pituitary adenoma [47]. Because this dataset was deposited in the

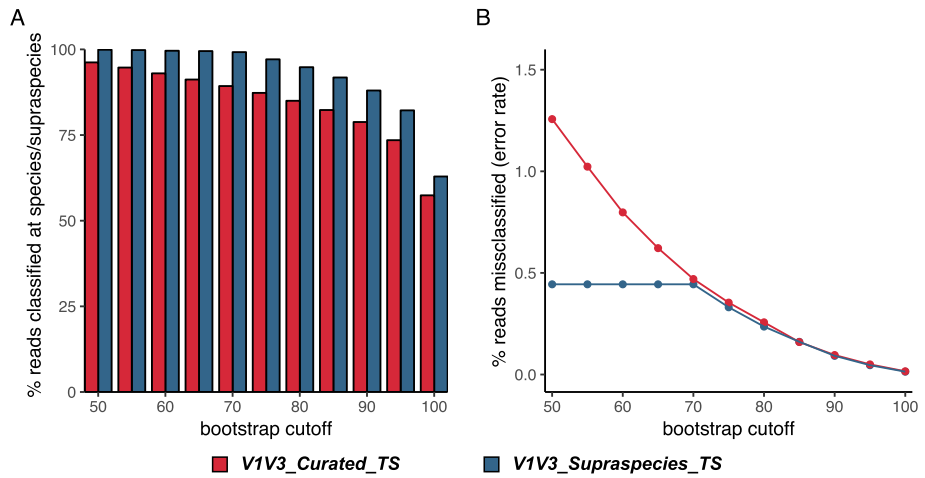


**Fig. 5** Addition of a supraspecies level to the training set increases the percentage of classified reads. **a** The percentage of eHOMD-derived simulated reads classified at species/supraspecies level using the V1V3_Curated_TS training set (red) versus the FL_Supraspecies_TS training set (blue). **b** The percentage of classified reads that were misclassified with each of these training sets. The naïve Bayesian RDP Classifier was used with bootstrap values ranging from 50 to 100

SRA, these sequences were not present in the NCBI nr/nt sequence repository (from which sequences were recruited for the eHOMD training set) and therefore not used in the generation of the eHOMD training set. We used the DADA2 pipeline for PacBio SMRT reads [48] to denoise these sequences and identify the relative abundance of each ASV in this dataset (FL_sinonasal_SMRT_ASV; Additional file 6). Next, we followed a three-pronged approach to assess the validity of the eHOMD training set for use with this fully independent sinonasal dataset.

First, we compared the best match by blastn, as a proxy for taxonomic assignment for these ASVs, using eHOMD compared to using NCBI 16S Microbial to identify any taxa in the dataset that might be missing in eHOMD, and thus the training set (Table 1; Additional file 7A). A detailed comparison (Additional file 7B) showed that 50 ASVs, corresponding to 11.1% of the total reads, were differentially assigned by blastn (98.5% identity cutoff) against eHOMDv15.1 versus against NCBI 16S Microbial. Of these 50 ASVs, 33 were assigned to taxa with a human microbial taxon (HMT) # as a provisional species designation. The majority (17/33) were assigned to *Peptoniphilus* sp. HMT-187, which corresponded in the NCBI 16S Microbial blastn output to *Peptoniphilus lacydonensis* and accounted for 7.6% of the total reads. Similarly, *Peptoniphilus* sp. HMT-187 was 11th in total reads across the nostril samples of 210 human microbiome project (HMP) participants [20]. The new species name *Peptoniphilus lacydonensis* was published in 2018 and is formally recognized [49]; therefore, it is now linked to HMT-187 in eHOMD. Of the remaining 17 ASVs, eight were mismatches between closely related *Streptococcus* species. Only six ASVs were nonassigned (NA) using eHOMD; thus, eHOMD included the vast majority (99.7%) of taxa present in the Earl-Mell sinonasal dataset. We note that based on recent additions, both eHOMD and NCBI 16S Microbial allowed correct assignment of ASVs to *Lawsonella clevelandensis*, which previously might be misidentified by blastn best match as a *Dietzia* species, and to *Neisseriaceae* G-1 bacterium HMT-327, which before receiving a stable taxonomic designation in eHOMD might be misidentified by blastn best match as a *Snodgrassella* species [20].

**Table 1** The eHOMD training sets performed well for species/supraspecies taxonomy assignment of an independent long-read 16S rRNA gene sinonasal dataset

| | # ASVs | # Reads | % ASVs | % Reads |
|---|---|---|---|---|
| NCBI 16S Microbial blastn | 178 | 140,455 | 87.3 | 97.3 |
| eHOMD blastn | 188 | 143,274 | 92.2 | 99.3 |
| eHOMD FL_Compilation_TS | 194 | 142,843 | 95.1 | 99.0 |
| eHOMD V1V3_Supraspecies_TS | 201 | 144,279 | 98.5 | 99.9 |

Columns three and four indicate the % of ASVs and reads, respectively, assigned subgenus-level taxonomy using each of the four approaches tested

Second, we compared species-level assignment of the dataset FL_sinonasal_SMRT_ASV using blastn against eHOMD versus the naïve Bayesian RDP Classifier with the full-length eHOMD training set FL_Compilation_TS. Less than 2% (1.91%) of the total reads in the dataset FL_sinonasal_SMRT_ASV, which are represented by 25 ASVs, were differentially assigned species-level taxonomy by these approaches (Table 1; Additional file 7A). There was no clear pattern to this differential assignment (Additional file 7C). This served to validate the use of the full-length eHOMD training set FL_

Compilation_TS and demonstrated that our approach is an effective method for assigning species-level taxonomy to close-to-full-length 16S rRNA gene sequences such as those derived from PacBio SMRT sequencing.

Third, based on the above, we used naïve Bayesian RDP Classifier-derived assignments with the training set FL_Compilation_TS as a proxy for the true species representation in the dataset FL_sinonasal_SMRT_ASV. We then used these to assess the performance of the training set V1V3_Supraspecies_TS for assigning species-level taxonomy to the V1–V3 region sequences of the dataset FL_sinonasal_SMRT_ASV (V1V3_sinonasal_SMRT_ASV; Additional file 8). There were 17 V1–V3 ASVs that were differentially assigned, corresponding to 1.64% of the reads (Table 1; Additional file 7D). Of these 17 differentially assigned ASVs, five belonged to a supraspecies, a level not represented in the full-length training set FL_Compilation_TS. Another four ASVs, all of which were present at very low relative abundance, had full-length ASVs assigned to *Anaerococcus tetradius*, whereas the V1–V3 region was assigned to *Anaerococcus octavius*, all with bootstrap values ≤ 92 (Additional file 7A). Therefore, when 16S rRNA gene sequences were resolved into ASVs based on their full-length and then cut to V1–V3, there was little to no species-level ambiguity (Additional file 7D). Thus, the ambiguity observed with V1–V3 16S rRNA short reads most likely occurs at the level of parsing into ASVs, due to fewer informative sequences, rather than at the assignment step with the naïve Bayesian RDP Classifier and the eHOMD training set.

### The eHOMD training set outperforms both the SILVA and RDP training sets

Having validated the use of the eHOMD training set with an independent 16S rRNA gene sinonasal dataset (Table 1), we next compared the use of our taxonomic assignment approach with other currently available pipelines that are coupled with the RDP or SILVA databases. We used three different datasets for this: (1) we generated a V1–V3 dataset derived from a collection of human aerodigestive tract 16S rRNA gene clone libraries (V1V3_hADT_CL; Additional file 9); (2) the HMP 16S rRNA gene V1–V3 454-sequenced nostril dataset that we previously analyzed [20]; and (3) the close-to-full-length ASVs of the FL_sinonasal_SMRT_ASV dataset. We then assigned genus-level taxonomy to all of these using the naïve Bayesian RDP Classifier with a bootstrap threshold of 70 coupled with three different training sets: an eHOMD training set (V1–V3 or FL), RDP16, or SILVA132 (the latter two from https://benjjneb.github.io/dada2/training.html). The eHOMD training set resulted in a larger percentage of reads assigned to a specific genus for each dataset; however, all three training sets resulted in genus-level assignment of > 90% of the sequences (Table 2). In contrast, a striking difference emerged between the different workflows when assigning taxonomy to species level. Our method of coupling the naïve Bayesian RDP Classifier with an eHOMD training set showed superior performance in terms of the percentage of reads classified compared to the exact string match method that is currently implemented in the DADA2 R Package in conjunction with SILVA132 or RDP16 (Table 2). As would be expected, the exact match algorithm assigned a higher percentage of close-to-full-length ASVs than V1–V3 region ASVs to taxonomy. The caveat being that these comprehensive databases have an estimated annotation error rates as high as ~ 10–17% [25]. For the V1–V3 region ASVs, it also performed much better with the V1–V3 sequences in the V1–V3 human aerodigestive tract clone library dataset than with the HMP V1–V3 dataset. We speculate that this occurred because the close-to-full-length sequences from the human aerodigestive tract clone library (V1V3_hADT_CL) dataset are part of both the RDP and SILVA databases, whereas the HMP V1–V3 454 sequences are not.

A key implication of these data is that our overall method yields comparable species-level results for 16S rRNA gene sequencing of the human aerodigestive tract using V1–V3 short-read sequences, which is very cost effective, compared to using close-to-full-length PacBio SMRT sequences. For example, it readily distinguished sequences of *Staphylococcus aureus* and *Staphylococcus epidermidis*. Another implication is that for species-level analysis of the microbiota of habitats lacking a high-resolution, accurate 16S rRNA gene database, PacBio SMRT sequencing coupled with the newly available DADA2 PacBio pipeline [48] to generate ASVs followed by blastn against NCBI 16S Microbial can provide effective species-level taxonomic assignment (Table 1; Additional file 7). Of note, the term supraspecies is not a valid taxonomic label, since it is dependent on the database and can vary for different short-read 16S rRNA gene regions. As such, a separate training set needs to be generated for each short-read region of interest. In addition, training sets need regular maintenance in conjunction with the database and need to be regenerated with each major revision of their associated database. In theory, other closed-reference methods for assigning taxonomy could benefit from the addition of an intermediate taxonomic level to preserve the highest level of resolution possible for that method, e.g., suprastrain between species and strain for metagenomics if there are strains that are too closely related to clearly distinguish.

## Conclusions

Here, we present a systematic approach for generating and validating habitat-specific 16S rRNA gene training

**Table 2** The eHOMD training set is superior for assigning species/supraspecies-level taxonomy to short- and long-read human aerodigestive tract datasets

| | | V1V3_hADT_CL | V1V3_HMPnares_ ASV[b] | | FL_sinonasal_SMRT_ASV | |
|---|---|---|---|---|---|---|
| | | (% reads) | (% ASVs) | (% reads) | (% ASVs) | (% reads) |
| eHOMD | Genus | 100.0 | 95.5 | 98.9 | 99.5 | 100.0 |
| | Species | 100.0 | 93.9 | 98.5 | 95.1 | 99.0 |
| SILVA | Genus | 96.1 | 94.7 | 97.6 | 96.6 | 98.9 |
| | Species | 44.7[a] | 4.1[a] | 29.9[a] | 18.6[a] | 71.9[a] |
| RDP | Genus | 93.2 | 90.2 | 92.2 | 94.1 | 98.5 |
| | Species | 38.5[a] | 3.1[a] | 27.5[a] | 13.2[a] | 60.6[a] |

[a]Exact match algorithm
[b]ASVs derived from the HMP nares V1–V3 dataset, as described in [20], constitute the V1V3_HMPnares_ASV dataset (Additional file 10)

sets to achieve species/supraspecies-level taxonomic assignment for short- or long-read 16S rRNA gene sequences. We used the naïve Bayesian RDP classifier with our training set; however, such a training set can be used with other taxonomic classifiers [41, 42, 50–57]. Our training set-construction approach includes several methodological advancements. First, we generated clusters of existing close-to-full length 16S rRNA gene sequences at the 99% level (with ≥ 98% coverage) around highly curated reference sequences. This collection of sequences for each taxon represented the currently known sequence variability for each taxon and increased the accuracy of taxonomic assignment using the naïve Bayesian RDP Classifier (Fig. 3). Second, we added an intermediate taxonomic level between genus and species, which we dubbed supraspecies, to which we assigned taxa in which sequences overlapped between two or more taxonomic clusters. This increased the percentage of sequences assigned subgenus-level taxonomy by preventing the default of consigning difficult-to-assign sequences to the genus level (Fig. 5). The process of constructing a short-read training set started with the generation of an effective close-to-full-length training set, which could then be used for taxonomic assignment of long-read 16S rRNA gene sequences such as those from clone libraries and from high-throughput long-read sequencing. As a final critical step, we validated the composition of the eHOMD training set using a PacBio SMRT-sequenced sinonasal dataset consisting of sequences independent of those used to build the training set (Table 1). Facilitating species/supraspecies-level analysis of 16S rRNA gene sequence datasets paves the way for cost-effective, population-scale molecular epidemiological microbiota studies that can achieve greater ecological/clinical relevance by reaching species level.

The successful implementation of a training set for species/supraspecies-level assignment of 16S rRNA gene short-read (e.g., V1–V3) datasets involved a collaborative give-and-take to iteratively optimize the sequencing protocol and the analysis workflow in conjunction with each other. As a result, here, we also provided three specific protocol recommendations for sequencing of the 16S rRNA gene V1–V3 region with the Illumina MiSeq. As detailed in Additional file 4, first, we show that excellent taxonomic assignment can be achieved with non-overlapping Illumina reads from V1 and V3 using our recommended sequencing protocol and taxonomic assignment workflow. This counters a common misperception that overlapping Illumina reads are required [6, 58]. The naïve Bayesian RDP Classifier tolerates nonoverlapping V1–V3 sequences; however, some other classifiers might not. Second, we demonstrated the value of initiating read 1 (R1) from the V3 reverse primer since R1 usually consists of a longer stretch of high-quality sequence. Third, we provide evidence for a role for increased amounts of PhiX when sequencing from the V1 primer, due to the high-degree of sequence conservation 3′ to this primer. This is analogous to the respective recommendations by Illumina and Mitra and colleagues to add PhiX to reach up to or, sometimes, greater than 50% of the total DNA for low diversity samples [59].

We note three limitations to the presented method for constructing a habitat-specific training set. First, capturing the natural sequence variation for the 16S rRNA gene(s) for each taxon is limited by the number of such sequences currently available in public repositories. For example, 140 of the ~ 770 taxa in eHOMD are represented by four or fewer distinct close-to-full-length sequences in the training set FL_Compilation_TS, whereas ~ 630 taxa are represented by five or greater. Second, the comprehensiveness of a training set depends on the comprehensiveness of the phylogeny-based, habitat-specific database from which it is constructed. Therefore, if an ASV is from a species that is absent from the database, then misclassification of that ASV to the most closely related species present in the database is possible. This can also occur during validation and benchmarking if using a simulated dataset with the same origin as the training set [43], e.g., our V1V3_eHOMDSim_250N100 dataset. This is a known limitation of closed reference-

based taxonomic assignment. As such, the known limitations of eHOMD are also limitations of any eHOMD training set [20, 26]. To overcome this limitation and increase accuracy for habitats that have not been sufficiently explored through sequencing, we recommend the use of a more conservative bootstrap value to reduce the error rate (Fig. 4c). Third, as with databases, training sets also require regular updating over time.

Within these limitations, we note several advantages. First, when coupled with a training set built with our method, the k-mer-based naïve Bayesian approach accommodates the natural variability of 16S rRNA gene sequences that exists within many bacterial species enabling high rates of accurate taxonomic assignment. In contrast, this natural variability limits the utility of any exact match algorithm to assigning species-level taxonomy for only those sequences already existing in a training set (Table 2). Second, despite all of the known limitations of a single-gene taxonomic indicator, the huge number of 16S rRNA gene sequences from diverse ecosystems available in public repositories supports the utility of the 16S rRNA gene for taxonomic assignment. In contrast, the utility of WGS metagenomic sequencing, which holds the promise of strain-level taxonomic assignment, remains limited by the quality and comprehensiveness of the genomic database used for closed-reference assignment. For example, at least one cultivar genome of each species is needed for more accurate species-level assignment. This remains problematic for habitats with many as-yet uncultivated species. Also, accurate strain-level assignment is dependent on the presence of cultivar genomes, and/or single-amplified genomes, of multiple strains of each species in the reference database. Further, a reference database should be free of chimeric metagenome-assembled genomes (MAGs) that combine genomic sequences that are unique to different strains of a species into one genome.

Finally, there are several additional recommendations to improve taxonomic assignment achieved with our method. First, it is critical to incorporate a stable provisional naming scheme into any habitat-specific database, e.g., the HMT numbers in eHOMD [20, 26]. Second, we recommend validating any training set generated with this method using a habitat-appropriate dataset that is fully independent of the sequences used to construct the training set, as we did with the sinonasal PacBio SMRT-generated dataset. Third, for ASVs that remain unassigned at finer taxonomic levels and are of interest based on their relative abundance and/or prevalence in the population, we recommend two additional steps: first, query these unassigned ASVs against a broad database such as SILVA, RDP, and/or NCBI 16S Microbial via blastn (≥ 98.5% identity over ≥ 98% coverage) to ascertain if these belong to a named species; and second, if no such named species match exists, then create a new provisional taxon in your database [20, 26]. This approach will identify candidate new taxa, both named and as-yet unnamed, for addition to a habitat-specific database and its associated training set/s.

## Methods

### Construction of eHOMD-based training datasets for the naïve Bayesian RDP Classifier

Training datasets were constructed in FASTA format as described on the DADA2 official website (https://benjjneb.github.io/dada2/index.html) [10], with taxonomy in the FASTA header line for each individual sequence. Training sets used in this study are described below.

#### FL_eHOMDrefs_TS

All of the full-length 16S eHOMDrefs, together with their respective taxonomic assignment, were formatted as a training set for the naïve Bayesian RDP Classifier, as above.

#### FL_Compilation_TS

This is the final version of the full-length 16S rRNA gene eHOMD training set, available as Additional file 2 and for download from eHOMD.org as eHOMDv15.1_FL_Compilation_TS.fa.gz. First, to more precisely calculate the percent identity for recruiting sequences for the training set, we trimmed each of the eHOMDrefs to nucleotides 28 and 1373. This is necessary for two reason: (1) several eHOMDrefs have hanging 5′ or 3′ ends that, if left in place, would affect the calculation of percent identity, and (2) this trimming permitted capture of sequences from the extensive datasets that use the reverse primer at ~ 1390. Second, each of the 998 trimmed eHOMDrefs was queried against the NCBI nonredundant nucleotide (nr/nt) database using blastn (NCBI BLAST 2.6.0+ package) (https://www.ncbi.nlm.nih.gov/books/NBK279690/) with the parameters -db nr -remote -perc_identity 97. Nucleotide sequences of GenBank IDs with ≥ 97% sequence identity to any of the eHOMDrefs were downloaded in FASTA format using the efetch.fcgi command of NCBI's Entrez Programming Utilities (E-utilities; https://www.ncbi.nlm.nih.gov/books/NBK25501/). The blastn hits were downloaded in the same orientation as the eHOMDrefs in two batches: (1) for subject sequence length between 1000 and 2000 nt, the entire subject sequence was downloaded; and (2) for subject sequence length > 2000 bp (e.g., a complete genomic sequence), only the aligned portion of the sequence was downloaded. Sequences < 1000 bp were not downloaded. Third, the 301,794 downloaded sequences, which matched to human microbial taxa (HMTs) at ≥ 97%, were parsed based on their highest sequence percent identity (≥ 99%) and alignment coverage (≥ 98% calculated based on the length of the reference) to any of

the eHOMDrefs in a given HMT. The choice of ≥ 99% identity was designed to obtain the centrally conserved set of sequences for each eHOMDref. The choice of ≥ 98% coverage was to ensure that the majority of the close-to-full-length sequence was present. Sequences that matched to multiple HMTs at equal percent identity and coverage were randomly assigned to only one HMT. (The use of supraspecies, see below, mitigates bioinformatic variability introduced at this step.) The FL_Compilation_TS training set is comprised of a total of 223,144 sequences parsed to their corresponding HMT, with a range of 1 to 4004 sequences per HMT. Additional file 11 is provided as a copy of the FL_Compilation_TS training set with sequences labeled as unique IDs, and Additional file 12 links those IDs to all the original GenBank accession numbers from which each training set sequence was derived.

#### V1V3_Raw_TS

We generated this training set version in two steps. First, each eHOMDref was individually aligned with the compilation of downloaded sequences that were matched to it at ≥ 99% identity and ≥ 98% coverage (see immediately above). Second, the sequences in the V1–V3 region, defined as positions 40-880 in the gapped eHOMDrefs alignment, were captured and then the alignment gaps removed. These steps were performed using a custom script (Additional file 13).

#### V1V3_Curated_TS

Sequences that were identical across V1–V3 in V1V3_Raw_TS were collapsed into a single sequence with the names of all taxa involved concatenated with a ":" separator. The majority of such concatenations occurred among either the same species, resulting in no name change, or different species of the same genus, resulting in assignment of a concatenated species name. However, there were a number of cases where species from two different genera were involved. These intergenus concatenations were carefully examined case by case. In all but one case, the concatenation was only supported by two to three sequences and, therefore, was deemed unreliable and rejected. After manual examination, only one intergenus concatenation remained. This was between HMT-559 (*Afipia broomeae*) and HMT-597 (*Bradyrhizobium elkanii*) and was supported by 29 sequences. The genus of the concatenated taxa was assigned as *Afipia*:*Bradyrhizobium* and species as *broomeae*:*elkanii*. Of note, although these two genera are almost identical on the V1–V3 region, they are 97% identical across the full length of the 16S rRNA gene.

#### V1V3_Supraspecies_TS

This is the final version of the V1–V3 16S rRNA gene eHOMD training set, available as Additional file 5 and for download from eHOMD.org as eHOMDv15.1_V1V3_Supraspecies_TS.fa.gz (http://www.homd.org/ftp/publication_data/20190709/). Additional file 14 is provided as a copy of the V1V3_Supraspecies_TS training set with sequences labeled as unique IDs, and Additional file 15 links those IDs to all the original GenBank accession numbers from which each training set sequence was derived. The sequences in this training set are the same as those included in the V1V3_Curated_TS version; only the header information was edited to include the supraspecies notation as a taxonomical level between genus and species. Therefore, instead of the header with seven levels included in previous versions of the eHOMD training set (i.e., >Kingdom;Phylum;Class;Order;Family;Genus;Species), the V1V3_Supraspecies_TS training set includes a header with eight levels (i.e., >Kingdom;Phylum;Class;Order;Family;Genus;Supraspecies;Species). A supraspecies was defined between two or more HMTs when there was a phylogenetic distance less than 0.005 between at least one pair of sequences of the corresponding HMTs. To calculate the distances, V1–V3 sequences were aligned with MAFTT v6.935b [60] and treed with FastTree version 2.1.9 [61] with default parameters. Pairwise distances between sequences were calculated as the sum of horizontal branch length between two sequence nodes based on FastTree's default "Jukes-Cantor + CAT" DNA evolution model [61]. The name of the resultant supraspecies is a concatenation of the species names of all HMTs involved, separated by ":" and is assigned at the supraspecies level (seventh level) for all the sequences of the involved HMTs. Sequences that were identical across the V1–V3 region between more than one taxon were assigned the concatenated name both at the supraspecies and species levels (e.g., >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;accolens:macginleyi:tuberculostearicum;accolens:macginleyi:tuberculostearicum). Whereas, the rest of the reads for those taxa were named at the supraspecies level as the concatenated name and maintained their unique identifier at the species level (e.g., >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;accolens:macginleyi:tuberculostearicum;accolens). For taxa not requiring concatenations (see the V1V3_Curated_TS description), the species designation is repeated at the supraspecies level such that the seventh- and eighth-level designations are identical (e.g., >Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Dolosigranulum;pigrum;pigrum). We note that the term supraspecies is dependent on the database/training set and, therefore, is not a valid formal taxonomic designation.

#### Generation of a V1–V3 16S rRNA gene-simulated dataset derived from eHOMD

Simulated reads were generated from each unique sequence in the V1V3_Curated_TS training set. During read generation, 1% of the bases in each unique initial sequence

were randomly selected and changed into a different base to simulate a 1% error rate. The resulting dataset contains 19,480 simulated sequences, each of them labeled with the ID of the parent eHOMDref sequence from which they were derived. Then, each error-simulated read was trimmed into two fragments to simulate Illumina pair-end reads (R1 starting from the V3 primer and R2 starting from the V1 primer). We generated multiple length configurations of both the R1 and R2 fragments. First, fixing R2 at 350 bp, we generated R1 fragments from the V3 primer ranging from 20 to 200 bp (Simulated_R2_350_R1_20-200, Additional file 16). Second, fixing R1 at 200 bp, we generated R2 fragments from the V1 primer ranging from 140 to 350 bp (Simulated_R2_140-350_R1_200, Additional file 17). Prior to analysis with the naïve Bayesian RDP Classifier, paired reads were connected with a "10N" linker to form one fragment ordered R2-10N-R1 (for further explanation, see Additional file 4). Subsequent evaluation of the steps for generating the eHOMD training sets in Figs. 3, 4, and 5 was done using the configuration R2(250 bp)-10N-R1(100 bp) (V1V3_eHOMDSim_250N100; Additional file 3).

### Reanalysis of a sinonasal PacBio-SMRT-sequenced full-length 16S rRNA gene dataset

Earl and colleagues kindly provided the circular consensus sequences (CCS) already demultiplexed, labeled by sample, pooled together and, then, filtered based on length distribution, terminal matches to the primer sequences, not aligning to a provided host or background genome sequence and with a cumulative expected error (EE < 1) [47]. We then used the DADA2 PacBio pipeline to denoise these CCS and identify the relative abundance of each ASV in the dataset; i.e., we used the learnErrors() function with errorEstimationFunction=PacBioErrfun and BAND_SIZE=32 [10, 48]. The sequences of the resulting 204 ASVs are in the FL_sinonasal_SMRT_ASV dataset (Additional file 6). The V1–V3 region was trimmed using bedtools getfasta with default parameters (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/) and reads were further trimmed into two fragments to simulate Illumina pair-end reads using the configuration R2(250 bp)-10N-R1(100 bp), as described above (V1V3_sinonasal_SMRT_ASV; Additional file 8).

### Generating a test V1–V3 human aerodigestive tract (hADT) microbiota dataset from full-length 16S rRNA gene clone libraries (CL)

We previously used the close-to-full-length 16S rRNA gene sequences from clone library-based microbiota studies of the human aerodigestive tract, as described in Supplemental Text S1 of [20]: Segre-Kong nostril (SKn) [62–67], Pei-Blaser [68, 69], Harris-Pace [70], van der Gast-Bruce [71], Flanagan-Bristow [72], and Perkins-Angenent [73]. Here, we compiled these into one dataset along with clones from NCBI PopSet UIDs 399192397, 399202217, 399199823, 399197584, 399194446, 399189902, 399186216, 399183739, 399182414, 399179617, 399175646, and 399173254 [74]. Aligned eHOMDrefs (eHOMDv15.1) sequences were trimmed from *Escherichia coli.* position 28-1373 and used to query this compiled dataset via blastn. We retained 27,816 sequences that hit with 100% coverage and ≥ 99.5% identity to 401 HMTs as the full-length human aerodigestivetract clone library dataset (FL_hADT_CL; Additional file 18). Of these, 5254 (18.9%) matched to more than one HMT, whereas 22,562 (81.1%) unambiguously matched to single HMTs. Sequences in this full-length CL dataset were then aligned using MAFTT v6.935b with default parameters [60]. Segments corresponding to the V1–V3 region were extracted based on positions of V1–V3 in the alignment (V1V3_hADT_CL, Additional file 9) using bedtools getfasta with default parameters (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/).

### Taxonomic assignment with the naïve Bayesian RDP Classifier

ASVs and CL sequences were assigned taxonomy using the naïve Bayesian RDP Classifier via the dada2::assignTaxonomy() function in R (an implementation of the algorithm with a k-mer size of 8 and 100 bootstrap iterations) with the specified training set and with outputBootstraps= TRUE and the indicated minBoot value [10, 39]. In addition to our eHOMD training sets, we used the RDP16 (*rdp_train_set_16.fa.gz*) and SILVA132 (*silva_nr_v132_train_set.fa.gz*) genus-level training sets available at https://benjjneb.github.io/dada2/training.html. Since the eHOMD training set V1V3_Supraspecies_TS generates an output with eight taxonomic levels that might not be compatible with common downstream applications that accept only seven levels, we provide a custom-written R function (Additional file 19) that converts the output from eight to seven levels as follows. First, if a sequence is not assigned at species level (i.e., the eighth level) at a given threshold (output as "NA"), then the eighth level is replaced with the classification result and bootstrap value from the corresponding seventh level. Next, the seventh level is deleted and only the resultant merged eighth level that can contain either species or supraspecies information is reported. Please note that this merged last supraspecies-or-species level is labeled simply as species.

### Taxonomic assignment with the DADA2 exact match

ASVs and CL sequences were assigned species-level taxonomy with the RDP16 (*rdp_species_assignment_16.fa.gz*) and SILVA132 (*silva_species_assignment_v132.fa.gz*) training set files downloaded from https://benjjneb.github.io/dada2/training.html using the dada2::assignSpecies() function in R with allowMultiple=TRUE [10].

### Taxonomic assignment with blastn

The NCBI BLAST 2.6.0+ package (https://www.ncbi.nlm.nih.gov/books/NBK279690/) was installed and the "blastn" command used with max_target_seqs 1, hits with < 98.5% identity were considered nonassigned. ASVs (FL_sinonasal_SMRT_ASV) derived from the sinonasal PacBio SMRT-sequenced dataset were also assigned taxonomy using blastn against two databases: the NCBI 16S Microbial database was downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/ on January 2019 [75] and eHOMDv15.1 was converted to a BLAST database using "makeblastdb" from the NCBI BLAST 2.6.0+ package.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s40168-020-00841-w.

**Additional file 1.** The expected effect of an uneven distribution of sequences among taxa in a training set for the Naïve Bayesian RDP Classifier.

**Additional file 2** eHOMDv15.1_FL_Compilation_TS.fa. The 16S rRNA gene full-length eHOMD training set (*FL_Compilation_TS*).

**Additional file 3.** V1V3_eHOMDSim_250N100.fa. The simulated eHOMD-derived V1V3_eHOMDSim_250N100 dataset.

**Additional file 4.** A method for achieving highly informative 16S rRNA gene V1-V3 region sequencing data using Illumina MiSeq.

**Additional file 5** eHOMDv15.1_V1V3_Supraspecies_TS.fa. The 16S rRNA gene V1-V3 eHOMD training set (***V1V3_Supraspecies_TS***).

**Additional file 6.** FL_sinonasal_SMRT_ASV.fa. Sequences of the resulting 204 ASVs in the Earl-Mell sinonasal 16S rRNA gene dataset (FL_sinonasal_SMRT_ASV dataset).

**Additional file 7.** Earl-Mell_sinonasal_analysis.xlsx with legend and tabs A-D. Species-level taxonomic assignment of the Earl-Mell sinonasal 16S rRNA gene dataset.

**Additional file 8.** V1V3_sinonasal_SMRT_ASV.fa. Sequences of the resulting 204 ASVs in the FL_sinonasal_SMRT_ASV dataset trimmed to the V1-V3 region (V1V3_sinonasal_SMRT_ASV dataset).

**Additional file 9** V1V3_hADT_CL.fa. V1-V3 trimmed 16S rRNA gene human aerodigestive tract clone library dataset (V1V3_hADT_CL dataset).

**Additional file 10.** V1V3_HMPnares_ASV.fa. ASVs derived from the HMP nares 16S rRNA gene V1-V3 dataset (V1V3_HMPnares_ASV dataset).

**Additional file 11.** *FL_Compilation_TS* training set with sequences labeled as unique IDs.

**Additional file 12.** Original GenBank accession numbers for all the sequences compiled in the *FL_Compilation_TS* training set linked to their unique IDs in Additional file 11.

**Additional file 13.** retrieveSegment.py. Custom script developed to generate a trimmed version of a training set (in our example to positions 40-880 in the gapped eHOMDrefs alignment) from a compilation of 16S rRNA gene full-length sequences.

**Additional file 14.** *V1V3_Supraspecies_TS* training set with sequences labelled as unique IDs.

**Additional file 15.** Original GenBank accession numbers for all the sequences compiled in the *V1V3_Supraspecies_TS* training set linked to their unique IDs in Additional file 14.

**Additional file 16.** Simulated_R2_350_R1_20-200.7z. Simulated eHOMD-derived dataset versions with R2 fixed at 350 bp and R1 fragments from the V3 primer ranging from 20 to 200 bp.

**Additional file 17.** Simulated_R2_140-350_R1_200.7z. Simulated eHOMD-derived dataset versions with R1 fixed at 200 bp and R2 fragments from the V1 primer ranging from 140 to 350 bp.

**Additional file 18.** FL_hADT_CL.fa. Full-length 16S rRNA gene human aerodigestive tract clone library dataset (FL_hADT_CL dataset).

**Additional file 19.** eight2seven.R. Custom-written R function that converts the dada2::assignTaxonomy() output from eight to seven taxonomic levels for compatibility with downstream applications.

### Acknowledgements

We are grateful to Joshua Earl, Joshua Chang Mell, and colleagues for providing the intermediate data table that facilitated analysis with a newly available algorithm and to members of the Lemon Lab and the Starr-Dewhirst-Johnston-Lemon Joint Group Meeting for helpful questions and suggestions throughout the project.

### Authors' contributions

YH, IFE, FED, and KPL conceived the project. YH, IFE, FED, and KPL designed the project. YH, IFE, TC, ML, and AK generated and analyzed the data. YH, IFE, TC, ML, AK, FED, and KPL interpreted the results. IFE and YH generated the figures and tables. KPL and IFE wrote the manuscript. All authors reviewed and edited the manuscript and approved the final version.

### Funding

This work was funded in part by a pilot grant (IFE, KPL) from the Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR001102 and financial contributions from Harvard University and its affiliated academic health care centers); by the National Institute of General Medical Sciences under award number R01GM117174 (KPL); by the National Institute of Allergy and Infectious Diseases under award number R01AI101018 (KPL); and by the National Institute of Dental and Craniofacial Research under award numbers R37DE016937 and R01DE024468 (FED). The content is solely the responsibility of the authors and does not reflect the official views of the National Institutes of Health or other funding source.

### Availability of data and materials

All data are included as additional files. The following are also available for download at eHOMD.org (http://www.homd.org/ftp/publication_data/20190709/): 1) the simulated eHOMD derived dataset (Additional file 3 as V1V3_eHOMDSim_250N100.fa), 2) both eHOMD training set files (Additional file 2 as eHOMDv15.1_FL_Compilation_TS.fa.gz; and Additional file 5 as eHOMTv15.1_V1V3_Supraspecies_TS.fa.gz) and 3) both custom scripts (Additional file 13 as retrieveSegment.py and Additional file 19 as eight2seven.R).

### Ethics approval and consent to participate

All participants provided informed consent and samples used to generate the V1–V3 region 16S rRNA dataset in Additional file 4 Figure B were collected under a protocol (#13-14 to FED) approved by the Forsyth Institutional Review Board.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Forsyth Institute (Microbiology), Cambridge, MA, USA. [2]Department of Oral Medicine, Infection & Immunity, Harvard School of Dental Medicine, Boston, MA, USA. [3]Department of Molecular Virology & Microbiology, Alkek Center for Metagenomics & Microbiome Research, Baylor College of Medicine, Houston, TX, USA. [4]Division of Infectious Diseases, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [5]Section of Infectious Diseases, Department of Pediatrics, Texas Children's Hospital and Baylor College of Medicine, Houston, TX, USA.

Received: 18 February 2020 Accepted: 15 April 2020
Published online: 15 May 2020

## References


1. Egerton S, Culloty S, Whooley J, Stanton C, Ross RP. The gut microbiota of marine fish. *Front Microbiol*. 2018;9:873.
2. Berendsen RL, van Verk MC, Stringlis IA, Zamioudis C, Tommassen J, Pieterse CM, Bakker PA. Unearthing the genomes of plant-beneficial Pseudomonas model strains WCS358, WCS374 and WCS417. *BMC Genomics*. 2015;16:539.
3. Brito IL, Alm EJ. Tracking strains in the microbiome: insights from metagenomics and models. *Front Microbiol*. 2016;7:712.
4. Mark Welch JL, Dewhirst FE, Borisy GG. Biogeography of the oral microbiome: the site-specialist hypothesis. *Annu Rev Microbiol*. 2019;73: 335–58.
5. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*. 2016;17:55.
6. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: attempting to find consensus "Best Practice" for 16S microbiome studies. *Appl Environ Microbiol*. 2018;84.
7. 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.
8. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in ecology and evolution*. 2013;4:1111–9.
9. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal*. 2015;9:968–79.
10. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13:581–3.
11. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 2017;11:2639–43.
12. Edgar RC: UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 2016:081257.
13. Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*. 2018;34:2371–5.
14. Tikhonov M, Leach RW, Wingreen NS. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J*. 2015;9:68–80.
15. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R: Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2017, 2.
16. Kumar PS, Brooker MR, Dowd SE, Camerlengo T. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS One*. 2011;6:e20956.
17. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*. 2016;17:135.
18. Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, Zhang H, Xiong Z, Xue Y, Tu J, Lu Z. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Sci Total Environ*. 2018;618:1254–67.
19. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*. 2019;6:190007.
20. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP: New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* 2018, 3.
21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41:D590–6.
22. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glockner FO. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*. 2014;42:D643–8.
23. 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:D633–42.
24. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. 2012;6:610–8.
25. Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ*. 2018;6:e5030.
26. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade WG. The human oral microbiome. *J Bacteriol*. 2010;192:5002–17.
27. Newton IL, Roeselers G. The effect of training set on the classification of honey bee gut microbiota using the Naive Bayesian Classifier. *BMC Microbiol*. 2012;12:221.
28. Seedorf H, Kittelmann S, Henderson G, Janssen PH. RIM-DB: a taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. *PeerJ*. 2014;2:e494.
29. Ritari J, Salojarvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics*. 2015;16:1056.
30. McIlroy SJ, Saunders AM, Albertsen M, Nierychlo M, McIlroy B, Hansen AA, Karst SM, Nielsen JL, Nielsen PH. MiDAS: the field guide to the microbes of activated sludge. Database (Oxford). 2015;2015:bav062.
31. Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR, Cole JR, Midgley DJ, Tran-Dinh N. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia*. 2016;108:1–5.
32. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, Albertsen M, Nielsen PH: MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. Database (Oxford) 2017, 2017.
33. Mikaelyan A, Kohler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. Classifying the bacterial gut microbiota of termites and cockroaches: a curated phylogenetic reference database (DictDb). *Syst Appl Microbiol*. 2015; 38:472–82.
34. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD: TaxAss: leveraging a custom freshwater database achieves fine-scale taxonomic resolution. *mSphere* 2018, 3.
35. Henderson G, Yilmaz P, Kumar S, Forster RJ, Kelly WJ, Leahy SC, Guan LL, Janssen PH. Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. *PeerJ*. 2019;7:e6496.
36. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE: The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* 2010, 2010:baq013.
37. Conlan S, Kong HH, Segre JA. Species-level analysis of DNA sequence data from the NIH Human Microbiome Project. *PLoS One*. 2012;7:e47075.
38. Al-Hebshi NN, Nasher AT, Idris AM, Chen T. Robust species taxonomy assignment algorithm for 16S rRNA NGS reads: application to oral carcinoma samples. *J Oral Microbiol*. 2015;7:28934.
39. 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;73:5261–7.
40. Lan Y, Wang Q, Cole JR, Rosen GL. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS One*. 2012;7:e32491.
41. Vinje H, Liland KH, Almoy T, Snipen L. Comparing K-mer based methods for improved classification of 16S sequences. *BMC Bioinformatics*. 2015;16:205.
42. Murali A, Bhargava A, Wright ES. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*. 2018;6:140.
43. Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*. 2018;6:e4652.
44. Russell S, Norvig P: Artificial Intelligence: A Modern Approach. Prentice Hall Press; 2009.
45. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J*. 2012;6:94–103.
46. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, Nadarajan R, Brodie EL, Lynch SV. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. *PLoS One*. 2015;10: e0117617.
47. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN, Workman AD, Blasetti M, Sen B, et al. Species-level bacterial community

profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome*. 2018;6:190.
48. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res*. 2019.
49. Beye M, Bakour S, Le Dault E, Rathored J, Michelle C, Cadoret F, Raoult D, Fournier PE. Peptoniphilus lacydonensis sp. nov., a new human-associated species isolated from a patient with chronic refractory sinusitis. *New Microbes New Infect*. 2018;23:61–9.
50. Nguyen NP, Mirarab S, Liu B, Pop M, Warnow T. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*. 2014;30:3548–55.
51. Allard G, Ryan FJ, Jeffery IB, Claesson MJ. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*. 2015;16:324.
52. Edgar RC: SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv* 2016:074161.
53. Gao X, Lin H, Revanna K, Dong Q. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*. 2017;18:247.
54. Liland KH, Vinje H. Snipen L: microclass: an R-package for 16S taxonomy classification. *BMC Bioinformatics*. 2017;18:172.
55. 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;33:3808–10.
56. 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;6:90.
57. Zheng Q, Bartow-McKenney C, Meisel JS, Grice EA. HmmUFOtu: An HMM and phylogenetic placement based ultra-fast taxonomic assignment and OTU picking tool for microbiome amplicon sequencing studies. *Genome Biol*. 2018;19:82.
58. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*. 2013;79:5112–20.
59. Mitra A, Skrzypczak M, Ginalski K, Rowicka M. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PLoS One*. 2015;10:e0120520.
60. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059–66.
61. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 2009;26:1641–50.
62. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, et al. Topographical and temporal diversity of the human skin microbiome. *Science*. 2009;324:1190–2.
63. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, Murray PR, et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome research*. 2012;22:850–9.
64. Oh J, Conlan S, Polley EC, Segre JA, Kong HH. Shifts in human skin and nares microbiota of healthy children and adults. *Genome medicine*. 2012;4:77.
65. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, Kong HH, Segre JA. Topographic diversity of fungal and bacterial communities in human skin. *Nature*. 2013;498:367–70.
66. Oh J, Freeman AF, Park M, Sokolic R, Candotti F, Holland SM, Segre JA, Kong HH. The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome research*. 2013;23:2103–14.
67. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. Biogeography and individuality shape function in the human skin metagenome. *Nature*. 2014;514:59–64.
68. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ. Bacterial biota in the human distal esophagus. *Proc Natl Acad Sci U S A*. 2004;101:4250–5.
69. Pei Z, Yang L, Peek RM, Jr Levine SM, Pride DT, Blaser MJ. Bacterial biota in reflux esophagitis and Barrett's esophagus. *World J Gastroenterol*. 2005;11:7277–83.
70. Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurso FJ, Pace NR. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. *Proc Natl Acad Sci U S A*. 2007;104:20529–33.
71. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW, Carroll MP, Parkhill J, Bruce KD. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J*. 2011;5:780–91.
72. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P, DeSantis TZ, Andersen GL, Wiener-Kronish JP, Bristow J. Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with Pseudomonas aeruginosa. *J Clin Microbiol*. 2007;45:1954–62.
73. Perkins SD, Woeltje KF, Angenent LT. Endotracheal tube biofilm inoculation of oral flora and subsequent colonization of opportunistic pathogens. *Int J Med Microbiol*. 2010;300:503–11.
74. Shelef KM: Ecology in the dentist's chair: patterns of biogeography and stability in human subgingival microbial communities. Stanford University, Department of Biology; 2013.
75. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733–45.

## Publisher's Note