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

Isabel F. Escapa[a,b,c,§] (Isabel.FernandezEscapa@bcm.edu), Yanmei Huang[a,b,§] (yhuang2@gmail.com), Tsute Chen[a,b] (tchen@forsyth.org), Maoxuan Lin[a] (mlin8@uncc.edu), Alexis Kokaras[a] (akorakas@forsyth.org), Floyd E. Dewhirst[a,b] (fdewhirst@forsyth.org), Katherine P. Lemon[a,c,d,e*] (Katherine.lemon@bcm.edu)

[a]Forsyth Institute (Microbiology), Cambridge, Massachusetts, USA

[b]Department of Oral Medicine, Infection & Immunity, Harvard School of Dental Medicine, Boston, Massachusetts, USA

[c]Alkek Center for Metagenomics & Microbiome Research, Department of Molecular Virology & Microbiology, Baylor College of Medicine, Houston, Texas, USA

[d]Division of Infectious Diseases, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[e]Section of Infectious Diseases, Department of Pediatrics, Texas Children's Hospital and Baylor College of Medicine, Houston, Texas, USA

[§] These authors contributed equally to this work.

*Correspondence: Katherine.lemon@bcm.edu

ABSTRACT

Background: The low cost of 16S rRNA gene sequencing facilitates population-scale molecular epidemiological studies. Existing computational algorithms can parse 16S rRNA gene sequences to 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 interested. 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 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.

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

61 distinct niches of a given environment [4]. High-throughput close-to-full-length 16S rRNA
62 gene sequencing (e.g., circular consensus sequences from PacBio Single Molecule,
63 Real-Time (SMRT) sequencing) and metagenomic whole genome sequencing (WGS)
64 hold promise for species- and strain-level microbiota studies. However, the easier
65 accessibility, and lower cost, of 16S rRNA gene short-read sequencing makes population-
66 scale (i.e., thousands of samples) molecular epidemiological studies of bacterial
67 microbiota of humans, other animals, plants, and the environment broadly feasible now.
68 A caveat to this is that the majority of published short-read 16S rRNA gene sequencing
69 studies use read clustering at a percent similarity that constrains resolution to the genus
70 level, i.e., 97% identity. Indeed, recent reviews on best practices and benchmarking for
71 16S rRNA gene microbiota studies focus on genus-level operational taxonomic unit
72 (OTU) analysis, e.g., [5-7]. However, newer algorithms, many of which are not based on
73 similarity thresholds, allow single-nucleotide resolution and can parse 16S rRNA gene
74 short-read sequences into species- or strain-level resolution phylotypes, usually called
75 amplicon sequence variants (ASVs) (e.g., MED (Minimal Entropy Decomposition) [8, 9],
76 DADA2 (Divisive Amplicon Denoising Algorithm) [10, 11] and UNOISE [12, 13], among
77 others [14, 15]). Another limitation of short-read 16S rRNA microbiome studies is that the
78 choice of the 16S rRNA gene region(s) sequenced places an upper bound on the degree
79 of species-level resolution that is achievable within a dataset [16-19]. Therefore, it is
80 critical to determine which regions provide the most information for distinguishing
81 bacterial species that are common to that specific ecosystem. For the habitats within the
82 human aerodigestive tract, i.e., the nasal passages, sinuses, throat, mouth, esophagus,
83 as well as the lower respiratory tract, we previously showed that many more taxa are

84 distinguishable at the species level with the V1-V3 region than with the commonly used
85 V3-V4 region of the 16S rRNA gene [20]. Therefore, we have developed a method that,
86 by combining the “reusability, reproducibility and comprehensiveness” of ASVs [11, 13]
87 and the selection of highly informative regions of the 16S rRNA gene, maximizes 16S
88 rRNA gene short-read sequencing potential to achieve sub-genus level resolution
89 taxonomic assignment.

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

107 for the human gut [29] and RIM-DP for rumen [28]. Second, custom addition of compiled
108 sequences from a specific habitat of interest can be used to augment a broad general
109 database, e.g., HBDB for honey bee [27], DictDB for termite and cockroach gut [33],
110 SILVA19Rum for rumen [35] and MiDAS for activated sludge [30, 32]. Third, both a
111 general and a habitat-specific database can be combined in the same pipeline, e.g., for
112 human-associated genera with pathogenic members [37], for freshwater (FreshTrain with
113 the TaxAss workflow [34]) and the multistage blastn workflow with eHOMD [38]. Many of
114 these databases are formatted for use as training sets to train classifier algorithms for
115 taxonomy assignment.

116 The naïve Bayesian RDP Classifier [39] is one of several effective algorithms for
117 assigning taxonomy. (For benchmarking and comparison with other methods, see [40-
118 43]). This type of supervised learning algorithm requires a training set, which is a set of
119 input-output examples to learn a function that can be used to make predictions [44]. In
120 this case, sequences are input and taxonomic assignments are output. Properly formatted
121 versions of the broad 16S rRNA gene databases SILVA, RDP and Greengenes are
122 available to train the most popular implementations of the naïve Bayesian RDP Classifier.
123 The quality of the training set strongly influences taxonomic assignment and habitat-
124 specific training sets have been developed to increase accuracy of taxonomic
125 assignments [27, 33, 40-42, 45]. However, the resolution of available training sets is
126 mostly limited to the genus level. An exception is the manually curated subset of the
127 Greengenes database corresponding to 89 clinically relevant bacterial genera that was
128 used to assign species-level taxonomy of full-length 16S rRNA gene sequences of clinical

129 isolates [46]. Notwithstanding, species-level taxonomy assignment of short-read 16S
130 rRNA gene datasets remains a challenge.

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

146 To test our hypothesis, we developed and validated short- and long-read training sets for
147 the microbiota of the human aerodigestive tract (mouth, nasal passages, sinuses, throat,
148 and esophagus) using our expanded Human Oral Microbiome Database (eHOMD). This
149 database was originally created and later expanded to serve as a resource for the
150 community of investigators generating datasets to study habitats within the human

151 aerodigestive tract [20, 26, 36]. In addition to 16S rRNA gene reference sequences
152 (eHOMDrefs), it also includes genomic and proteomic data. (It also works well for the
153 lower respiratory tract [20].) The lack of proper taxonomical representation in traditional
154 databases is a challenge in predicting taxonomic assignments [43]. A strength of the
155 eHOMD is that, by placing 16S rRNA gene reference sequences for each Human
156 Microbial Taxon (HMT) on a phylogenetic tree
157 (http://ehomd.org/index.php?name=HOMD&show_tree=_), as-yet unnamed or
158 uncultivated species are defined based on sequence identity and added to the phylogeny
159 using a provisional naming scheme that permits taxonomic assignment for cross-study
160 comparison [26]. Also, sequences that are misnamed in other databases are easily
161 identified and given a correct designation in eHOMD. Another key strength of eHOMD is
162 that it is locally comprehensive often allowing approximately 95% of sequences from V1-
163 V3 aerodigestive tract datasets to be assigned to the taxonomy [20].

164 **RESULTS.**

165 **Compiling closely-related sequences for each taxon in a training set improves the**
166 **accuracy of species-level taxonomic classification.** Genus-level taxonomic
167 assignment is not an inherent limitation of the naïve Bayesian RDP Classifier. Rather,
168 taxonomic assignment to 16S rRNA gene short reads is limited by both the resolution to
169 which sequences in datasets are parsed and by the nature of the training set used. The
170 former is overcome by using approaches such as oligotyping/MED [8, 9], DADA2 [10, 11]
171 or unoise2 [12, 25] to parse sequence variants at high resolution. We hypothesized that
172 the limitations inherent in training sets could also be overcome.

173 The naïve Bayesian RDP Classifier algorithm indicates that a training set with a larger
174 number of sequences representing each taxon will result in more confident taxonomic
175 assignment [39]. Based on the conditional probability for a member of a taxon (T), the
176 higher the occurrence of a given distinguishing "k-mer" (word or w_i) in the training set, the
177 greater the confidence with which assignment of that taxon is made, i.e., more sequences
178 can be classified unambiguously. Thus, as the number of sequences (M) for each taxon
179 in the training set increases (**Figure 2A vs. 2B**), the number of accurate assignments
180 should increase. (See **Additional File 1** for a more detailed explanation.) In support of
181 this, Werner and colleagues found more sequence counts result in more assignments
182 [45]. Therefore, to systematically increase M, we used each of the reference sequences
183 in eHOMD, i.e., the eHOMDrefs, as bait to capture closely matching, publicly available
184 sequences and combined the resulting compilation of sequences for each taxon into a
185 close-to-full-length compilation training set (***FL_Compilation_TS***) that reflects the
186 currently known 16S rRNA gene sequence variability for each taxon, both natural and
187 sequencing-error derived (**Additional File 2**).

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

196 that classified at the species-level at incremental bootstrap values from 50 to 100 using
197 the training set ***FL_Compilation_TS*** (**Figure 3A,** orange bars) compared with the training
198 set ***FL_eHOMDrefs_TS*** (**Figure 3A,** purple bars), which consists of only the eHOMDrefs.
199 Doing this, we observed an increase in the percentage of reads that classified at species-
200 level with the compilation TS except at a bootstrap value of 100. We postulate that the
201 additional sequences classified with the ***FL_eHOMDrefs_TS*** TS at a bootstrap of 100 are
202 misclassified sequences. A higher rate of misclassification is expected with the training
203 set ***FL_eHOMDrefs_TS***, since it includes only a few representative sequences for each
204 taxon. Therefore, we next assayed for any misclassifications. At each bootstrap threshold,
205 the percentage of reads that were misclassified (i.e., reads for which the assigned
206 taxonomic identity was different than the known identity of the original sequence from
207 which the simulated read was derived) was at least 50% lower using the training set
208 ***FL_Compilation_TS*** (**Figure 3B**, orange line) than with the training set
209 ***FL_eHOMDrefs_TS*** (**Figure 3B**, purple line). Thus, classification of the dataset
210 **V1V3_eHOMDSim_250N100** showed a reduced error rate and increased confidence
211 level when using a training set consisting of a compilation of closely related sequences
212 rather than one consisting of only one or a few reference sequences for each taxon.

213 **Moving towards an appropriate short-read fragment training set.** The training set
214 ***FL_Compilation_TS*** consisted of close-to-full-length 16S rRNA gene sequences. We
215 hypothesized that the presence of k-mers outside the V1-V3 region in the training set
216 might lead to misclassifications when using it with a V1-V3 region dataset. Supporting
217 this, when using a training set based on a large general database, trimming reference
218 sequences to match the sequenced region increases the number of sequences that are

219 assigned taxonomy [45]. Therefore, we trimmed the sequences in the training set
220 ***FL_Compilation_TS*** to cover only the V1-V3 region generating training set
221 ***V1V3_Raw_TS*** (**Figure 2C**). Using this there was no gain in the percentage of reads in
222 the simulated dataset **V1V3_eHOMDSim_250N100** classified to the species-level
223 (**Figure 4A**, green bar). Moreover, we observed an increase in the percentage of
224 misclassified reads (**Figure 4B**, green line), i.e. the assignment accuracy decreased.
225 Therefore, we next determined why the use of appropriate short-read fragments in this
226 training set paradoxically decreased confidence and accuracy of species-level
227 assignment and resolved the issue.

228 **Combining closely related, indistinguishable taxa into supraspecies decreases the**
229 **error rate for a short-read training set.** Considering the possible explanations for the
230 above paradox, we realized that taxa with distinct full-length 16S rRNA gene sequences
231 can have identical V1-V3 sequences. In silico, using only V1-V3, 37 of the ~770 species-
232 level taxa in eHOMD are no longer distinguishable from at least one other species at
233 100% identity [20]. Therefore, we hypothesized that the observed loss in accuracy using
234 training set ***V1V3 _Raw_TS*** was due to identical sequences with more than one species
235 name, e.g., *Veillonella parvula* and *Veillonella dispar* [20]. To solve this problem, we
236 removed duplicate sequences and assigned a combined name, i.e., a supraspecies name
237 such as *Veillonella* parvula_dispar, to the remaining unique sequence (**Figure 2D**). This
238 resulted in the training set ***V1V3_Curated_TS***, which showed improved accuracy
239 compared with both the ***FL_Compilation_TS*** and the ***V1V3_Raw_TS*** (**Figure 4B**, red
240 line). However, this improvement came at a cost of a 0.7 to 4.4 % decrease in the reads
241 assigned supraspecies- or species-level taxonomy at each bootstrap threshold (**Figure**

242 **4A**, red bar). This trade-off can be illustrated by graphing at each bootstrap threshold the
243 percentage of reads from the simulated **V1V3_eHOMDSim_250N100** dataset that were
244 misclassified versus confidently assigned species-level taxonomy using the
245 ***V1V3_Curated_TS*** with the naïve Bayesian RDP Classifier (**Figure 4C**).

246 **Maximizing performance of the naïve Bayesian RDP Classifier for subgenus-level**
247 **taxonomic assignment of short-read sequences requires both insertion of**
248 **supraspecies as a taxonomic level and setting a threshold bootstrap**. Two steps
249 were required to reap the benefits of using the supraspecies definition. First, formally
250 inserting supraspecies as a taxonomic level between genus and species in the name
251 header for each sequence in the training set, which yielded the training set
252 ***V1V3_Supraspecies_TS*** (**Figure 2E; Additional File 5**). Second, establishing the
253 bootstrap cut-off at which a sequence is not assigned at the species level so that the
254 naïve Bayesian RDP Classifier will then default to the supraspecies level, rather than
255 defaulting to genus, allowing for a higher percentage of reads to be assigned (**Figure 5A**,
256 blue bar). This latter step preserves subgenus level information encoded in the ASVs.
257 This choice inevitably involves a trade-off between accuracy and percentage of reads
258 classified below genus level, e.g., see **Figure 4C**. For our purposes, we chose a
259 conservative bootstrap value of 70 (**Figure 5B,** blue line). With the simulated data, for
260 which the truth is known, this gave an error rate of around 0.05%. Of note, although one
261 common bootstrap setting for the naïve Bayesian RDP Classifier is 50, we use of a more
262 conservative value for species-level taxonomic assignment with the
263 ***V1V3_Supraspecies_TS***.

264 **Using an independent sinonasal dataset validated the performance of the eHOMD**
265 **training set.** We next validated the performance of the eHOMD training set against a
266 dataset of close-to-full-length 16S rRNA gene sequences from sinonasal samples. Very
267 importantly, these sequences are independent of the sequences recruited to build the
268 training set. Earl and colleagues performed PacBio SMRT sequencing of close-to-full-
269 length 16S rRNA genes from multiple sinonasal samples from each of 12 adults without
270 sinus inflammation who were undergoing removal of a pituitary adenoma [47]. Because
271 this dataset was deposited in the SRA, these sequences were not present in the NCBI
272 nr/nt sequence repository (from which sequences were recruited for the eHOMD training
273 set) and therefore not used in the generation of the eHOMD training set. We used the
274 DADA2 pipeline for PacBio SMRT reads [48] to denoise these sequences and identify the
275 relative abundance of each ASV in this dataset (**FL_sinonasal_SMRT_ASV; Additional**
276 **File 6**). Next, we followed a three-pronged approach to assess the validity of the eHOMD
277 training set for use with this fully independent sinonasal dataset.

278 First, we compared the best match by blastn, as a proxy for taxonomic assignment for
279 these ASVs, using eHOMD compared to using NCBI 16S Microbial to identify any taxa in
280 the dataset that might be missing in eHOMD, and thus the training set (**Table 1;**
281 **Additional File 7A**). A detailed comparison (**Additional File 7B**), showed that 50 ASVs,
282 corresponding to 11.1% of the total reads, were differentially assigned by blastn (98.5%
283 identity cutoff) against eHOMDv15.1 versus against NCBI 16S Microbial. Of these 50
284 ASVs, 33 were assigned to taxa with an HMT (Human Microbial Taxon) # as a provisional
285 species designation. The majority (17/33) were assigned to *Peptoniphilus* sp. HMT-187,
286 which corresponded in the NCBI 16S Microbial blastn output to *Peptoniphilus*

287 *lacydonensis* and accounted for 7.6 % of the total reads. Similarly, *Peptoniphilus* sp.
288 HMT-187 was 11th in total reads across the nostril samples of 210 HMP (Human
289 Microbiome Project) participants [20]. The new species name *Peptoniphilus lacydonensis*
290 was published in 2018 and is formally recognized [49]; therefore, it is now linked to HMT-
291 187 in eHOMD. Of the remaining 17 ASVs, 8 were mismatches between closely related
292 *Streptococcus* species. Only 6 ASVs were nonassigned (NA) using eHOMD, thus,
293 eHOMD included the vast majority (99.7%) of taxa present in the Earl-Mell sinonasal
294 dataset. We note that based on recent additions both eHOMD and NCBI 16S Microbial
295 allowed correct assignment of ASVs to *Lawsonella clevelandensis*, which previously
296 might be misidentified by blastn best match as a Dietzia species, and to *Neisseriaceae*
297 G-1 bacterium HMT-327, which before receiving a stable taxonomic designation in
298 eHOMD might be misidentified by blastn best match as a *Snodgrassella* species [20].

299 Second, we compared species-level assignment of the dataset
300 **FL_sinonasal_SMRT_ASV** using blastn against eHOMD versus the naïve Bayesian
301 RDP Classifier with the full-length eHOMD training set ***FL_Compilation_TS***. Less than
302 2% (1.91%) of the total reads in the dataset **FL_sinonasal_SMRT_ASV**, which are
303 represented by 25 ASVs, were differentially assigned species-level taxonomy by these
304 approaches (**Table 1; Additional File 7A**). There was no clear pattern to this differential
305 assignment (**Additional File 7C**). This served to validate the use of the full-length
306 eHOMD training set ***FL_Compilation_TS*** and demonstrated that our approach is an
307 effective method for assigning species-level taxonomy to close-to-full-length 16S rRNA
308 gene sequences such as those derived from PacBio SMRT sequencing.

309 Third, based on the above, we used naïve Bayesian RDP Classifier-derived assignments
310 with the training set ***FL_Compilation_TS*** as a proxy for the true species representation
311 in the dataset **FL_sinonasal_SMRT_ASV**. We then used these to assess the
312 performance of the training set ***V1V3_Supraspecies_TS*** for assigning species-level
313 taxonomy to the V1-V3 region sequences of the dataset **FL_sinonasal_SMRT_ASV**
314 (**V1V3_sinonasal_SMRT_ASV; Additional File 8**). There were seventeen V1-V3 ASVs
315 that were differentially assigned, corresponding to 1.64% of the reads (**Table 1;**
316 **Additional File 7D**). Of these 17 differentially-assigned ASVs, 5 belonged to a
317 supraspecies, a level not represented in the full-length training set ***FL_Compilation_TS***.
318 Another 4 ASVs, all of which were present at very low relative abundance, had full-length
319 ASVs assigned to *Anaerococcus tetradius*, whereas the V1-V3 region was assigned to
320 *Anaerococcus octavius*, all with bootstrap values ≤ 92 (**Additional File 7A**). In contrast,
321 when 16S rRNA gene sequences were parsed into ASVs based on their full-length and
322 then cut to V1-V3, there was little to no species-level ambiguity (**Additional File 7D**).
323 Thus, the ambiguity observed with V1-V3 16S rRNA short reads most likely occurs at the
324 level of parsing into ASVs, due to fewer informative sequences, rather than at the
325 assignment step with the naïve Bayesian RDP Classifier and the eHOMD training set.

326 **Table 1. The eHOMD training sets performed well for species/supraspecies**
327 **taxonomy assignment of an independent long-read 16S rRNA gene sinonasal**
328 **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

Table 1 Legend. Columns three and four indicate the % of ASVs and reads, respectively, assigned subgenus-level taxonomy using each of the four approaches tested.

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

351 assigned a higher percentage of close-to-full-length ASVs than V1-V3 region ASVs to
352 taxonomy. The caveat being that these comprehensive databases have an estimated
353 annotation error rates as high as ~10-17% [25]. For the V1-V3 region ASVs, it also
354 performed much better with the V1-V3 sequences in the V1-V3 human aerodigestive tract
355 clone library dataset than with the HMP V1-V3 dataset. We speculate this occurred
356 because the close-to-full-length sequences from the human aerodigestive tract clone
357 library (**V1V3_hADT_CL**) dataset are part of both the RDP and SILVA databases,
358 whereas the HMP V1-V3 454 sequences are not.

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

		V1V3_hADT_CL	**V1V3_HMPnares_ ASV#**		**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*	4.1*	29.9*	18.6*	71.9*
RDP	Genus	93.2	90.2	92.2	94.1	98.5
	Species	38.5*	3.1*	27.5*	13.2*	60.6*

361 **Table 2 Legend**. *Exact match algorithm; # ASVs derived from the HMP nares V1-V3
362 dataset, as described in [20], constitute the **V1V3_HMPnares_ASV** dataset (**Additional**
363 **File 10**).

364 A key implication of these data is that our overall method yields comparable species-level
365 results for 16S rRNA gene sequencing of the human aerodigestive tract using V1-V3
366 short-read sequences, which is very cost effective, compared to using close-to-full-length
367 PacBio SMRT sequences. For example, it readily distinguished sequences of

368 *Staphylococcus aureus* and *Staphylococcus epidermidis*. Another implication is that for
369 species-level analysis of the microbiota of habitats lacking a high-resolution, accurate
370 16S rRNA gene database, PacBio SMRT sequencing coupled with the newly available
371 DADA2 PacBio pipeline [48] to generate ASVs followed by blastn against NCBI 16S
372 Microbial can provide effective species-level taxonomic assignment (**Table 1**). Of note,
373 the definition of supraspecies is dependent on the database and can vary for different
374 short-read 16S rRNA gene regions. As such, a separate training set needs to be
375 generated for each short-read region of interest. In addition, training sets need regular
376 maintenance in conjunction with the database and need to be regenerated with each
377 major revision of their associated database. In theory, other closed-reference methods
378 for assigning taxonomy could benefit from the addition of an intermediate taxonomic level
379 to preserve the highest level of resolution possible for that method, e.g., suprastrain
380 between species and strain for metagenomics if there are strains that are too closely
381 related to clearly distinguish.

382 **CONCLUSIONS.**

383 Here, we present a systematic approach for generating and validating habitat-specific
384 16S rRNA gene training sets to achieve species/supraspecies-level taxonomic
385 assignment for short- or long-read 16S rRNA gene sequences. We used the naïve
386 Bayesian RDP classifier with our training set; however, such a training set can be used
387 with other taxonomic classifiers [41, 42, 50-57]. Our training set-construction approach
388 includes several methodological advancements. First, we generated clusters of existing
389 close-to-full length 16S rRNA gene sequences at the 99% level (with ≥ 98% coverage)

390 around highly curated reference sequences. This collection of sequences for each taxon
391 represented the currently known sequence variability for each taxon and increased the
392 accuracy of taxonomic assignment using the naïve Bayesian RDP Classifier (**Figure 3**).
393 Second, we added an intermediate taxonomic level between genus and species, which
394 we dubbed supraspecies, to which we assigned taxa in which sequences overlapped
395 between two or more taxonomic clusters. This increased the percentage of sequences
396 assigned subgenus-level taxonomy by preventing the default of consigning difficult-to-
397 assign sequences to the genus level (**Figure 5**). The process of constructing a short-read
398 training set started with the generation of an effective close-to-full-length training set,
399 which could then be used for taxonomic assignment of long-read 16S rRNA gene
400 sequences such as those from clone libraries and from high-throughput long-read
401 sequencing. As a final critical step, we validated the composition of the eHOMD training
402 set using a PacBio SMRT-sequenced sinonasal dataset consisting of sequences
403 independent of those used to build the training set (**Table 1**). Facilitating
404 species/supraspecies-level analysis of 16S rRNA gene sequence datasets paves the way
405 for cost-effective, population-scale molecular epidemiological microbiota studies that can
406 achieve greater ecological/clinical relevance by reaching species level.

407 The successful implementation of a training set for species/supraspecies-level
408 assignment of 16S rRNA gene short-read (e.g., V1-V3) datasets involved a collaborative
409 give-and-take to iteratively optimize the sequencing protocol and the analysis workflow in
410 conjunction with each other. As a result, here, we also provided three specific protocol
411 recommendations for sequencing of the 16S rRNA gene V1-V3 region with the Illumina
412 MiSeq. As detailed in **Additional File 4**, first, we show that excellent taxonomic

413 assignment can be achieved with nonoverlapping Illumina reads from V1 and V3 using
414 our recommended sequencing protocol and taxonomic assignment workflow. This
415 counters a common misperception that overlapping Illumina reads are required [6, 58].
416 The naïve Bayesian RDP Classifier tolerates nonoverlapping V1-V3 sequences;
417 however, some other classifiers might not. Second, we demonstrated the value of
418 initiating Read 1 (R1) from the V3 reverse primer since R1 usually consists of a longer
419 stretch of high-quality sequence. Third, we provide evidence for a role for increased
420 amounts of PhiX when sequencing from the V1 primer, due to the high-degree of
421 sequence conservation 3' to this primer. This is analogous to the respective
422 recommendations by Illumina and Mitra and colleagues to add PhiX to reach up to or,
423 sometimes, greater than 50% of the total DNA for low diversity samples [59].

424 We note four limitations to the presented method for constructing a habitat-specific
425 training set. First, capturing the natural sequence variation for the 16S rRNA gene(s) for
426 each taxon is limited by the number of such sequences currently available in public
427 repositories. For example, 140 of the ~770 taxa in eHOMD are represented by four or
428 fewer distinct close-to-full-length sequences in the training set ***FL_Compilation_TS***,
429 whereas ~630 taxa are represented by five or greater. Second, the comprehensiveness
430 of a training set depends on the comprehensiveness of the phylogeny-based, habitat-
431 specific database from which it is constructed. Therefore, if an ASV is from a species that
432 is absent from the database, then misclassification of that ASV to the mostly closely
433 related species present in the database is possible. This can also occur during validation
434 and benchmarking if using a simulated dataset with the same origin as the training set
435 [43], e.g., our **V1V3_eHOMDSim_250N100** dataset. This is a known limitation of closed

436 reference-based taxonomic assignment. As such, the known limitations of eHOMD are
437 also limitations of any eHOMD training set [20, 26]. Third, as with databases, training sets
438 also require regular updating over time.

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

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

471 **METHODS.**

472 **Construction of eHOMD-based training datasets for the naïve Bayesian RDP**
473 **Classifier.** Training datasets were constructed in FASTA format as described on the
474 DADA2 official website (https://benjjneb.github.io/dada2/index.html) [10], with taxonomy
475 in the FASTA header line for each individual sequence. Training sets used in this study
476 are described below.

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

480 ***FL_Compilation_TS***. This is the final version of the full-length 16S rRNA gene eHOMD
481 training set, available as **Additional File 2** and for download from eHOMD.org as
482 ***eHOMDv15.1_FL_Compilation_TS.fa.gz***. First, to more precisely calculate the percent
483 identity for recruiting sequences for the training set, we trimmed each of the eHOMDrefs
484 to nucleotides 28 and 1373. This is necessary for two reason: 1) several eHOMDrefs have
485 hanging 5' or 3' ends that, if left in place, would affect the calculation of percent identity,
486 and 2) this trimming permitted capture of sequences from the extensive datasets that use
487 the reverse primer at ~1390. Second, each of the 998 trimmed eHOMDrefs was queried
488 against the NCBI nonredundant nucleotide (nr/nt) database using blastn (NCBI BLAST
489 2.6.0+ package) (https://www.ncbi.nlm.nih.gov/books/NBK279690/) with the parameters
490 -db nr -remote -perc_identity 97. Nucleotide sequences of GenBank IDs with ≥ 97%
491 sequence identity to any of the eHOMDrefs were downloaded in FASTA format using the
492 efetch.fcgi command of NCBI's Entrez Programming Utilities (E-utilities;
493 https://www.ncbi.nlm.nih.gov/books/NBK25501/). The blastn hits were downloaded in the
494 same orientation as the eHOMDrefs in two batches: 1) for subject sequence length
495 between 1000 and 2000 nt, the entire subject sequence was downloaded; and 2) for
496 subject sequence length > 2000 bp (e.g., a complete genomic sequence), only the aligned
497 portion of the sequence was downloaded. Sequences < 1000 bp were not downloaded.
498 Third, the 301,794 downloaded sequences, which matched to Human Microbial Taxa
499 (HMTs) at ≥ 97%, were parsed based on their highest sequence percent identity (≥ 99%)

500 and alignment coverage (≥ 98% calculated based on the length of the reference) to any
501 of the eHOMDrefs in a given HMT. The choice of ≥ 99% identity was designed to obtain
502 the centrally conserved set of sequences for each eHOMDref. The choice of ≥ 98%
503 coverage was to ensure that the majority of the close-to-full-length sequence was present.
504 Sequences that matched to multiple HMTs at equal percent identity and coverage were
505 randomly assigned to only one HMT. (The use of supraspecies, see below, mitigates
506 bioinformatic variability introduced at this step.) The ***FL_Compilation_TS*** training set is
507 comprised of a total of 223,144 sequences parsed to their corresponding HMT, with a
508 range of 1 to 4004 sequences per HMT.

509 ***V1V3_Raw_TS***. We generated this training set version in two steps. First, each
510 eHOMDref was individually aligned with the compilation of downloaded sequences that
511 were matched to it at ≥ 99% identity and ≥ 98% coverage (see immediately above).
512 Second, the sequences in the V1-V3 region, defined as positions 40-880 in the gapped
513 eHOMDrefs alignment, were captured and then the alignment gaps removed. These steps
514 were performed using a custom script (**Additional File 11**).

515 ***V1V3_Curated_TS***. Sequences that were identical across V1-V3 in ***V1V3_Raw_TS*** were
516 collapsed into a single sequence with the names of all taxa involved concatenated with a
517 ":" separator. The majority of such concatenations occurred among either the same
518 species, resulting in no name change, or different species of the same genus, resulting
519 in assignment of a concatenated species name. However, there were a number of cases
520 where species from two different genera were involved. These intergenus concatenations
521 were carefully examined case by case. In all but one case, the concatenation was only

522 supported by two to three sequences and, therefore, was deemed unreliable and rejected.
523 After manual examination, only one intergenus concatenation remained. This was
524 between HMT-559 (*Afipia broomeae*) and HMT-597 (*Bradyrhizobium elkanii*) and was
525 supported by 29 sequences. The genus of the concatenated taxa was assigned as
526 *Afipia:Bradyrhizobium* and species as *broomeae:elkanii*. Of note, although these two
527 genera are almost identical on the V1-V3 region, they are 97% identical across the full
528 length of the 16S rRNA gene.

529 **V1V3_Supraspecies_TS**. This is the final version of the V1-V3 16S rRNA gene eHOMD
530 training set, available as **Additional File 5** and for download from eHOMD.org as
531 eHOMDv15.1_V1V3_Supraspecies_TS.fa.gz
532 (http://www.homd.org/ftp/publication_data/20190709/). The sequences in this training set
533 are the same as those included in the ***V1V3_Curated_TS*** version; only the header
534 information was edited to include the supraspecies notation as a taxonomical level
535 between genus and species. Therefore, instead of the header with seven-levels included
536 in previous versions of the eHOMD training set (i.e.,
537 >Kingdom;Phylum;Class;Order;Family;Genus;Species), the ***V1V3_Supraspecies_TS***
538 training set includes a header with eight levels (i.e.,
539 >Kingdom;Phylum;Class;Order;Family;Genus;Supraspecies;Species). A supraspecies
540 was defined between two or more HMTs when there was a phylogenetic distance less
541 than 0.005 between at least one pair of sequences of the corresponding HMTs. To
542 calculate the distances, V1V3 sequences were aligned with MAFTT v6.935b [60] and
543 treed with FastTree version 2.1.9 [61] with default parameters. Pairwise distances
544 between sequences were calculated as the sum of horizontal branch length between two

545 sequence nodes based on FastTree's default "Jukes-Cantor + CAT" DNA evolution model
546 [61]. The name of the resultant supraspecies is a concatenation of the species names of
547 all HMTs involved, separated by ":" and is assigned at the supraspecies level (seventh
548 level) for all the sequences of the involved HMTs. Sequences that were identical across
549 the V1-V3 region between more than one taxon were assigned the concatenated name
550 both at the supraspecies and species levels (e.g.,
551 >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Coryne-
552 bacterium;accolens:macginleyi:tuberculostearicum;accolens:macginleyi:tuberculostearic
553 um). Whereas, the rest of the reads for those taxa were named at the supraspecies level
554 as the concatenated name and maintained their unique identifier at the species level (e.g.,
555 >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Coryne-
556 bacterium;accolens:macginleyi:tuberculostearicum;accolens). For taxa not requiring
557 concatenations (see the ***V1V3_Curated_TS*** description), the species designation is
558 repeated at the supraspecies level such that the seventh- and eighth-level designations
559 are identical (e.g., >Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Dolo-
560 sigranulum;pigrum;pigrum).

561 **Generation of a V1-V3 16S rRNA gene simulated dataset derived from eHOMD.**
562 Simulated reads were generated from each unique sequence in the ***V1V3_Curated_TS***
563 training set. During read generation, one percent of the bases in each unique initial
564 sequence were randomly selected and changed into a different base to simulate a 1%
565 error rate. The resulting dataset contains 19,480 simulated sequences, each of them
566 labelled with the ID of the parent eHOMDref sequence from which they were derived.
567 Then, each error-simulated read was trimmed into two fragments to simulate Illumina pair-

568 end reads (R1 starting from the V3 primer and R2 starting from the V1 primer). We
569 generated multiple length configurations of both the R1 and R2 fragments. First, fixing R2
570 at 350 bp, we generated R1 fragments from the V3 primer ranging from 20 to 200 bp
571 (Simulated_R2_350_R1_20-200, **Additional File 12**). Second, fixing R1 at 200 bp, we
572 generated R2 fragments from the V1 primer ranging from 140 to 350 bp
573 (Simulated_R2_140-350_R1_200, **Additional File 13**). Prior to analysis with the naïve
574 Bayesian RDP Classifier, paired reads were connected with a "10N" linker to form one
575 fragment ordered R2-10N-R1 (For further explanation see **Additional File 4**).
576 Subsequent evaluation of the steps for generating the eHOMD training sets in Figures 3
577 to 5 was done using the configuration R2(250 bp)-10N-R1(100 bp)
578 **(V1V3_eHOMDSim_250N100; Additional File 3**).

579 **Reanalysis of a sinonasal PacBio-SMRT-sequenced full-length 16S rRNA gene**
580 **dataset.** Earl and colleagues kindly provided the circular consensus sequences (CCS)
581 already demultiplexed, labeled by sample, pooled together and, then, filtered based on
582 length distribution, terminal matches to the primer sequences, not aligning to a provided
583 host or background genome sequence and with a cumulative expected error (EE < 1)
584 [47]. We then used the DADA2 PacBio pipeline to denoise these CCS and identify the
585 relative abundance of each ASV in the dataset; i.e., we used the learnErrors() function
586 with errorEstimationFunction=PacBioErrfun and BAND_SIZE=32 [10, 48]. The
587 sequences of the resulting 204 ASVs are in the **FL_sinonasal_SMRT_ASV** dataset
588 (**Additional File 6**). The V1-V3 region was trimmed using bedtools getfasta with default
589 parameters (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/) and
590 reads were further trimmed into two fragments to simulate Illumina pair-end reads using

591 the configuration R2(250 bp)-10N-R1(100 bp), as described above
592 (**V1V3_sinonasal_SMRT_ASV**; **Additional File 8**).

593 **Generating a test V1-V3 human aerodigestive tract (hADT) microbiota dataset from**
594 **full-length 16S rRNA gene clone libraries (CL).** We previously used the close-to-full-
595 length 16S rRNA gene sequences from clone library-based microbiota studies of the
596 human aerodigestive tract, as described in Supplemental Text S1 of [20]: Segre-Kong
597 nostril (SKn) [62-67], Pei-Blaser [68, 69], Harris-Pace [70], van der Gast-Bruce [71],
598 Flanagan-Bristow [72] and Perkins-Angenent [73]. Here, we compiled these into one
599 dataset along with clones from NCBI PopSet UIDs 399192397, 399202217, 399199823,
600 399197584, 399194446, 399189902, 399186216, 399183739, 399182414, 399179617,
601 399175646 and 399173254 [74]. Aligned eHOMDrefs (eHOMDv15.1) sequences were
602 trimmed from *E. coli.* position 28-1373 and used to query this compiled dataset via blastn.
603 We retained 27,816 sequences that hit with 100% coverage and ≥ 99.5% identity to 401
604 HMTs as the full-length human aerodigestive tract clone library dataset (**FL_hADT_CL;**
605 **Additional File 14**). Of these, 5254 (18.9%) matched to more than one HMT, whereas
606 22,562 (81.1%) unambiguously matched to single HMTs. Sequences in this full-length CL
607 dataset were then aligned using MAFTT v6.935b with default parameters [60]. Segments
608 corresponding to the V1-V3 region were extracted based on positions of V1-V3 in the
609 alignment (**V1V3_hADT_CL**, **Additional File 9**) using bedtools getfasta with default
610 parameters (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/).

611 **Taxonomic assignment with the naïve Bayesian RDP Classifier.** ASVs and CL
612 sequences were assigned taxonomy using the naïve Bayesian RDP Classifier via the

613 dada2::assignTaxonomy() function in R (an implementation of the algorithm with a k-mer
614 size of 8 and 100 bootstrap iterations) with the specified training set and with
615 outputBootstraps=TRUE and the indicated minBoot value [10, 39]. In addition to our
616 eHOMD training sets, we used the RDP16 (***rdp_train_set_16.fa.gz***) and SILVA132
617 (***silva_nr_v132_train_set.fa.gz***) genus-level training sets available at
618 https://benjjneb.github.io/dada2/training.html. Since the eHOMD training set
619 ***V1V3_Supraspecies_TS*** generates an output with eight taxonomic levels that might not
620 be compatible with common downstream applications that accept only seven levels, we
621 provide a custom-written R function (**Additional File 15**) that converts the output from
622 eight to seven levels as follows. First, if a sequence is not assigned at species level (i.e.,
623 the eighth level) at a given threshold (output as "NA") then the eighth level is replaced
624 with the classification result and bootstrap value from the corresponding seventh level.
625 Next, the seventh level is deleted and only the resultant merged eighth level that can
626 contain either species or supraspecies information is reported. Please note that this
627 merged last supraspecies-or-species level is labeled simply as species.

628 **Taxonomic assignment with the DADA2 exact match.** ASVs and CL sequences were
629 assigned species-level taxonomy with the RDP16 (***rdp_species_assignment_16.fa.gz***)
630 and SILVA132 (***silva_species_assignment_v132.fa.gz***) training set files downloaded
631 from https://benjjneb.github.io/dada2/training.html using the dada2::assignSpecies()
632 function in R with allowMultiple=TRUE [10].

633 **Taxonomic assignment with blastn.** The NCBI BLAST 2.6.0+ package
634 (https://www.ncbi.nlm.nih.gov/books/NBK279690/) was installed and the "blastn"

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

641 **DECLARATIONS.**

642 **Ethics approval and consent to participate.** All participants provided informed consent
643 and samples used to generate the V1-V3 region 16S rRNA dataset in **Additional File 4**
644 **Figure B** were collected under a protocol (#13-14 to FED) approved by the Forsyth
645 Institutional Review Board**.**

646 **Consent for publication.** Not applicable.

647 **Availability of data and material.** All data are included as additional files. The following
648 are also available for download at eHOMD.org
649 (http://www.homd.org/ftp/publication_data/20190709/): 1) the simulated eHOMD derived
650 dataset (**Additional File 3** as V1V3_eHOMDSim_250N100.fa), 2) both eHOMD training
651 set files (**Additional File 2** as eHOMDv15.1_FL_Compilation_TS.fa.gz; and **Additional**
652 **File 5** as eHOMTv15.1_V1V3_Supraspecies_TS.fa.gz) and 3) both custom scripts
653 (**Additional File 11** as retrieveSegment.py and **Additional File 15** as eight2seven.R).

654 **Competing interests.** The authors declare that they have no competing interests.

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.

Authors' contributions. Conceived Project: YH, IFE, FED, KPL. Designed Project: YH, IFE, FED, KPL. Generated and Analyzed data: YH, IFE, TC, ML, AK. Interpreted results: YH, IFE, TC, ML, AK, FED, KPL. Generated figures and tables: IFE, YH. Wrote manuscript: KPL, IFE. All authors reviewed and edited the manuscript and approved the final version.

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.

676 **REFERENCES**

677 1. Egerton S, Culloty S, Whooley J, Stanton C, Ross RP: **The Gut Microbiota of**
678 **Marine Fish.** *Front Microbiol* 2018, **9:**873.
679 2. Berendsen RL, van Verk MC, Stringlis IA, Zamioudis C, Tommassen J, Pieterse
680 CM, Bakker PA: **Unearthing the genomes of plant-beneficial Pseudomonas**
681 **model strains WCS358, WCS374 and WCS417.** *BMC Genomics* 2015, **16:**539.
682 3. Brito IL, Alm EJ: **Tracking Strains in the Microbiome: Insights from**
683 **Metagenomics and Models.** *Front Microbiol* 2016, **7:**712.
684 4. Welch JLM, Dewhirst FE, Borisy GG: **Biogeography of the Oral Microbiome:**
685 **The Site-Specialist Hypothesis.** *Annu Rev Microbiol* 2019.
686 5. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M,
687 Podar M, Quince C, Hall N: **A comprehensive benchmarking study of**
688 **protocols and sequencing platforms for 16S rRNA community profiling.** *BMC*
689 *Genomics* 2016, **17:**55.
690 6. Pollock J, Glendinning L, Wisedchanwet T, Watson M: **The Madness of**
691 **Microbiome: Attempting To Find Consensus "Best Practice" for 16S**
692 **Microbiome Studies.** *Appl Environ Microbiol* 2018, **84**.
693 7. Almeida A, Mitchell AL, Tarkowska A, Finn RD: **Benchmarking taxonomic**
694 **assignments based on 16S rRNA gene profiling of the microbiota from**
695 **commonly sampled environments.** *Gigascience* 2018, **7**.
696 8. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML:
697 **Oligotyping: Differentiating between closely related microbial taxa using 16S**

698 **rRNA gene data.** *Methods in ecology and evolution / British Ecological Society*
699 2013, **4**.
700 9. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML: **Minimum**
701 **entropy decomposition: unsupervised oligotyping for sensitive partitioning**
702 **of high-throughput marker gene sequences.** *The ISME journal* 2015, **9:**968-
703 979.
704 10. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP: **DADA2:**
705 **High-resolution sample inference from Illumina amplicon data.** *Nat Methods*
706 2016, **13:**581-583.
707 11. Callahan BJ, McMurdie PJ, Holmes SP: **Exact sequence variants should**
708 **replace operational taxonomic units in marker-gene data analysis.** *ISME J*
709 2017, **11:**2639-2643.
710 12. Edgar RC: **UNOISE2: improved error-correction for Illumina 16S and ITS**
711 **amplicon sequencing.** *bioRxiv* 2016**:**081257.
712 13. Edgar RC: **Updating the 97% identity threshold for 16S ribosomal RNA OTUs.**
713 *Bioinformatics* 2018, **34:**2371-2375.
714 14. Tikhonov M, Leach RW, Wingreen NS: **Interpreting 16S metagenomic data**
715 **without clustering to achieve sub-OTU resolution.** *ISME J* 2015, **9:**68-80.
716 15. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z,
717 Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R: **Deblur Rapidly**
718 **Resolves Single-Nucleotide Community Sequence Patterns.** *mSystems* 2017,
719 **2**.

720 16. Kumar PS, Brooker MR, Dowd SE, Camerlengo T: **Target region selection is a**
721 **critical determinant of community fingerprints generated by 16S**
722 **pyrosequencing.** *PLoS One* 2011, **6:**e20956.
723 17. Yang B, Wang Y, Qian P-Y: **Sensitivity and correlation of hypervariable**
724 **regions in 16S rRNA genes in phylogenetic analysis.** *BMC bioinformatics* 2016,
725 **17:**135-135.
726 18. Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, Zhang H, Xiong Z, Xue Y, Tu J, Lu
727 Z: **Evaluation of different 16S rRNA gene V regions for exploring bacterial**
728 **diversity in a eutrophic freshwater lake.** *Sci Total Environ* 2018, **618:**1254-
729 1267.
730 19. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI:
731 **The effect of 16S rRNA region choice on bacterial community metabarcoding**
732 **results.** *Scientific Data* 2019, **6:**190007.
733 20. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP: **New Insights**
734 **into Human Nostril Microbiome from the Expanded Human Oral Microbiome**
735 **Database (eHOMD): a Resource for the Microbiome of the Human**
736 **Aerodigestive Tract.** *mSystems* 2018, **3**.
737 21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner
738 FO: **The SILVA ribosomal RNA gene database project: improved data**
739 **processing and web-based tools.** *Nucleic Acids Res* 2013, **41:**D590-596.
740 22. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies
741 J, Ludwig W, Glockner FO: **The SILVA and "All-species Living Tree Project**
742 **(LTP)" taxonomic frameworks.** *Nucleic Acids Res* 2014, **42:**D643-648.

743 23. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro
744 A, Kuske CR, Tiedje JM: **Ribosomal Database Project: data and tools for high**
745 **throughput rRNA analysis.** *Nucleic Acids Res* 2014, **42:**D633-642.
746 24. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A,
747 Andersen GL, Knight R, Hugenholtz P: **An improved Greengenes taxonomy**
748 **with explicit ranks for ecological and evolutionary analyses of bacteria and**
749 **archaea.** *ISME J* 2012, **6:**610-618.
750 25. Edgar R: **Taxonomy annotation and guide tree errors in 16S rRNA databases.**
751 *PeerJ* 2018, **6:**e5030.
752 26. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A,
753 Wade WG: **The human oral microbiome.** *J Bacteriol* 2010, **192:**5002-5017.
754 27. Newton IL, Roeselers G: **The effect of training set on the classification of**
755 **honey bee gut microbiota using the Naive Bayesian Classifier.** *BMC Microbiol*
756 2012, **12:**221.
757 28. Seedorf H, Kittelmann S, Henderson G, Janssen PH: **RIM-DB: a taxonomic**
758 **framework for community structure analysis of methanogenic archaea from**
759 **the rumen and other intestinal environments.** *PeerJ* 2014, **2:**e494.
760 29. Ritari J, Salojarvi J, Lahti L, de Vos WM: **Improved taxonomic assignment of**
761 **human intestinal 16S rRNA sequences by a dedicated reference database.**
762 *BMC Genomics* 2015, **16:**1056.
763 30. McIlroy SJ, Saunders AM, Albertsen M, Nierychlo M, McIlroy B, Hansen AA, Karst
764 SM, Nielsen JL, Nielsen PH: **MiDAS: the field guide to the microbes of**
765 **activated sludge.** *Database (Oxford)* 2015, **2015:**bav062.

766 31. Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR,
767 Cole JR, Midgley DJ, Tran-Dinh N: **Fungal identification using a Bayesian**
768 **classifier and the Warcup training set of internal transcribed spacer**
769 **sequences.** *Mycologia* 2016, **108:**1-5.
770 32. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM,
771 Albertsen M, Nielsen PH: **MiDAS 2.0: an ecosystem-specific taxonomy and**
772 **online database for the organisms of wastewater treatment systems**
773 **expanded for anaerobic digester groups.** *Database (Oxford)* 2017, **2017**.
774 33. Mikaelyan A, Kohler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A:
775 **Classifying the bacterial gut microbiota of termites and cockroaches: A**
776 **curated phylogenetic reference database (DictDb).** *Syst Appl Microbiol* 2015,
777 **38:**472-482.
778 34. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD: **TaxAss: Leveraging a**
779 **Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution.**
780 *mSphere* 2018, **3**.
781 35. Henderson G, Yilmaz P, Kumar S, Forster RJ, Kelly WJ, Leahy SC, Guan LL,
782 Janssen PH: **Improved taxonomic assignment of rumen bacterial 16S rRNA**
783 **sequences using a revised SILVA taxonomic framework.** *PeerJ* 2019,
784 **7:**e6496.
785 36. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE: **The Human**
786 **Oral Microbiome Database: a web accessible resource for investigating oral**
787 **microbe taxonomic and genomic information.** *Database (Oxford)* 2010,
788 **2010:**baq013.

789 37. Conlan S, Kong HH, Segre JA: **Species-level analysis of DNA sequence data**
790 **from the NIH Human Microbiome Project.** *PLoS One* 2012, **7:**e47075.
791 38. Al-Hebshi NN, Nasher AT, Idris AM, Chen T: **Robust species taxonomy**
792 **assignment algorithm for 16S rRNA NGS reads: application to oral**
793 **carcinoma samples.** *J Oral Microbiol* 2015, **7:**28934.
794 39. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid**
795 **assignment of rRNA sequences into the new bacterial taxonomy.** *Appl*
796 *Environ Microbiol* 2007, **73:**5261-5267.
797 40. Lan Y, Wang Q, Cole JR, Rosen GL: **Using the RDP classifier to predict**
798 **taxonomic novelty and reduce the search space for finding novel organisms.**
799 *PLoS One* 2012, **7:**e32491.
800 41. Vinje H, Liland KH, Almoy T, Snipen L: **Comparing K-mer based methods for**
801 **improved classification of 16S sequences.** *BMC Bioinformatics* 2015, **16:**205.
802 42. Murali A, Bhargava A, Wright ES: **IDTAXA: a novel approach for accurate**
803 **taxonomic classification of microbiome sequences.** *Microbiome* 2018, **6:**140.
804 43. Edgar RC: **Accuracy of taxonomy prediction for 16S rRNA and fungal ITS**
805 **sequences.** *PeerJ* 2018, **6:**e4652.
806 44. Russell S, Norvig P: *Artificial Intelligence: A Modern Approach.* Prentice Hall
807 Press; 2009.
808 45. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG,
809 Angenent LT, Knight R, Ley RE: **Impact of training sets on classification of**
810 **high-throughput bacterial 16s rRNA gene surveys.** *ISME J* 2012, **6:**94-103.

811 46. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S,
812 Nadarajan R, Brodie EL, Lynch SV: **Use of 16S rRNA gene for identification of**
813 **a broad range of clinically relevant bacterial pathogens.** *PLoS One* 2015,
814 **10:**e0117617.
815 47. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN,
816 Workman AD, Blasetti M, Sen B, et al: **Species-level bacterial community**
817 **profiling of the healthy sinonasal microbiome using Pacific Biosciences**
818 **sequencing of full-length 16S rRNA genes.** *Microbiome* 2018, **6:**190.
819 48. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK,
820 Dougherty MK: **High-throughput amplicon sequencing of the full-length 16S**
821 **rRNA gene with single-nucleotide resolution.** *Nucleic Acids Res* 2019.
822 49. Beye M, Bakour S, Le Dault E, Rathored J, Michelle C, Cadoret F, Raoult D,
823 Fournier PE: **Peptoniphilus lacydonensis sp. nov., a new human-associated**
824 **species isolated from a patient with chronic refractory sinusitis.** *New*
825 *Microbes New Infect* 2018, **23:**61-69.
826 50. Nguyen NP, Mirarab S, Liu B, Pop M, Warnow T: **TIPP: taxonomic identification**
827 **and phylogenetic profiling.** *Bioinformatics* 2014, **30:**3548-3555.
828 51. Allard G, Ryan FJ, Jeffery IB, Claesson MJ: **SPINGO: a rapid species-classifier**
829 **for microbial amplicon sequences.** *BMC Bioinformatics* 2015, **16:**324.
830 52. Edgar RC: **SINTAX: a simple non-Bayesian taxonomy classifier for 16S and**
831 **ITS sequences.** *bioRxiv* 2016**:**074161.

832 53. Gao X, Lin H, Revanna K, Dong Q: **A Bayesian taxonomic classification**
833 **method for 16S rRNA gene sequences with improved species-level accuracy.**
834 *BMC Bioinformatics* 2017, **18:**247.
835 54. Liland KH, Vinje H, Snipen L: **microclass: an R-package for 16S taxonomy**
836 **classification.** *BMC Bioinformatics* 2017, **18:**172.
837 55. Matias Rodrigues JF, Schmidt TSB, Tackmann J, von Mering C: **MAPseq: highly**
838 **efficient k-mer search with confidence estimates, for rRNA sequence**
839 **analysis.** *Bioinformatics* 2017, **33:**3808-3810.
840 56. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA,
841 Gregory Caporaso J: **Optimizing taxonomic classification of marker-gene**
842 **amplicon sequences with QIIME 2's q2-feature-classifier plugin.** *Microbiome*
843 2018, **6:**90.
844 57. Zheng Q, Bartow-McKenney C, Meisel JS, Grice EA: **HmmUFOtu: An HMM and**
845 **phylogenetic placement based ultra-fast taxonomic assignment and OTU**
846 **picking tool for microbiome amplicon sequencing studies.** *Genome Biol* 2018,
847 **19:**82.
848 58. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD: **Development of**
849 **a dual-index sequencing strategy and curation pipeline for analyzing**
850 **amplicon sequence data on the MiSeq Illumina sequencing platform.** *Appl*
851 *Environ Microbiol* 2013, **79:**5112-5120.
852 59. Mitra A, Skrzypczak M, Ginalski K, Rowicka M: **Strategies for achieving high**
853 **sequencing accuracy for low diversity samples and avoiding sample**
854 **bleeding using illumina platform.** *PLoS One* 2015, **10:**e0120520.

855 60. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid**
856 **multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids*
857 *Res* 2002, **30:**3059-3066.
858 61. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution**
859 **trees with profiles instead of a distance matrix.** *Mol Biol Evol* 2009, **26:**1641-
860 1650.
861 62. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG,
862 Blakesley RW, Murray PR, Green ED, et al: **Topographical and temporal**
863 **diversity of the human skin microbiome.** *Science* 2009, **324:**1190-1192.
864 63. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley
865 EC, Komarow HD, Murray PR, et al: **Temporal shifts in the skin microbiome**
866 **associated with disease flares and treatment in children with atopic**
867 **dermatitis.** *Genome research* 2012, **22:**850-859.
868 64. Oh J, Conlan S, Polley EC, Segre JA, Kong HH: **Shifts in human skin and nares**
869 **microbiota of healthy children and adults.** *Genome medicine* 2012, **4:**77.
870 65. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos
871 E, Park M, Kong HH, Segre JA: **Topographic diversity of fungal and bacterial**
872 **communities in human skin.** *Nature* 2013, **498:**367-370.
873 66. Oh J, Freeman AF, Park M, Sokolic R, Candotti F, Holland SM, Segre JA, Kong
874 HH: **The altered landscape of the human skin microbiome in patients with**
875 **primary immunodeficiencies.** *Genome research* 2013, **23:**2103-2114.

876 67. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA: **Biogeography and**
877 **individuality shape function in the human skin metagenome.** *Nature* 2014,
878 **514:**59-64.
879 68. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ: **Bacterial biota in the**
880 **human distal esophagus.** *Proc Natl Acad Sci U S A* 2004, **101:**4250-4255.
881 69. Pei Z, Yang L, Peek RM, Jr Levine SM, Pride DT, Blaser MJ: **Bacterial biota in**
882 **reflux esophagitis and Barrett's esophagus.** *World J Gastroenterol* 2005,
883 **11:**7277-7283.
884 70. Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess
885 H, Deterding RR, Accurso FJ, Pace NR: **Molecular identification of bacteria in**
886 **bronchoalveolar lavage fluid from children with cystic fibrosis.** *Proc Natl*
887 *Acad Sci U S A* 2007, **104:**20529-20533.
888 71. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW,
889 Carroll MP, Parkhill J, Bruce KD: **Partitioning core and satellite taxa from within**
890 **cystic fibrosis lung bacterial communities.** *ISME J* 2011, **5:**780-791.
891 72. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P,
892 DeSantis TZ, Andersen GL, Wiener-Kronish JP, Bristow J: **Loss of bacterial**
893 **diversity during antibiotic treatment of intubated patients colonized with**
894 **Pseudomonas aeruginosa.** *J Clin Microbiol* 2007, **45:**1954-1962.
895 73. Perkins SD, Woeltje KF, Angenent LT: **Endotracheal tube biofilm inoculation of**
896 **oral flora and subsequent colonization of opportunistic pathogens.** *Int J Med*
897 *Microbiol* 2010, **300:**503-511.

898 74. Shelef KM: **Ecology in the dentist's chair: patterns of biogeography and**
899 **stability in human subgingival microbial communities.** Stanford University,
900 Department of Biology; 2013.
901 75. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,
902 Robbertse B, Smith-White B, Ako-Adjei D, et al: **Reference sequence (RefSeq)**
903 **database at NCBI: current status, taxonomic expansion, and functional**
904 **annotation.** *Nucleic Acids Res* 2016, **44:**D733-745.

905

906

907 **FIGURE LEGENDS.**

908 **Figure 1. Relationships between the datasets, databases and training sets in**
909 **constructing training sets for a specific habitat: the human aerodigestive tract. A)**
910 Datasets gathered from public repositories or obtained by sequencing of new samples
911 are used to explore the 16S rRNA gene diversity of the habitat of interest. These include
912 both 16S rRNA full-length sequences and region-specific short-read sequences used for
913 method validation or benchmarking. **B)** A curated habitat-specific full-length 16S rRNA
914 gene reference database is assembled and expanded in an iterative way by selecting
915 from those datasets in A representative sequences for both named and as-yet unnamed
916 or uncultivated species (i.e., HMTs in eHOMD), and placing them in a phylogenetic tree
917 (See Figure 1 in [20]). **C)** Training sets are derived from the taxonomical hierarchy of the
918 habitat-specific database and enhanced by the following steps: compiling multiple 16S
919 rRNA gene sequences to represent the natural sequence variability of each taxon,
920 trimming the training set to match the sequenced region/s, and placing species sharing
921 closely related sequences into a supraspecies taxonomic level. To enhance clarity, the
922 ***names of training sets*** appear in bold italics whereas the **names of datasets** appear in
923 bold throughout this paper. Datasets in grey are the specific examples used for the
924 construction of the eHOMD derived training sets described here. Solid arrows indicate
925 where the sequences described come from and dotted arrows indicate when datasets
926 were used for validation or benchmarking.

927 **Figure 2. Schematic representation of the steps to generate sequential habitat-**
928 **specific training sets.** (**a**) The ***FL_eHOMDrefs_TS*** training set contains all full-length

929 eHOMDrefs (thick lines) from eHOMDv15.1 together with their respective taxonomic
930 assignment. When only one read represents each taxon (M=1) a given distinguishing k-
931 mer (green fragment) can only be either present (1) or absent (0). (**b**) A higher number of
932 sequences per taxon (M) allows for better resolution on the assignment, with the presence
933 of a given distinguishing k-mer (green fragment) across each cluster of reads (w_i) being
934 represented as a proportion out of the total number of reads in that taxon (M). Therefore,
935 to better represent the known sequence diversity of the 16S rRNA gene(s) for each taxon,
936 the training set ***FL_Compilation_TS*** includes clusters of sequences (thin lines)
937 recovered from the NCBI nonredundant nucleotide (nr/nt) database that matched with
938 99% identity and ≥ 98% coverage (see methods) to each eHOMDref (thick line). (**c**) The
939 training set ***V1V3_Raw_TS*** is a V1-V3 trimmed version of the ***FL_Compilation_TS***
940 training set. The schematic illustrates how trimming to this region leads to identical reads
941 (purple lines) having two different taxonomic designations. Here, G is genus and species
942 are labeled as A or B. (**d**) To construct the ***V1V3_Curated_TS*** training set, identical V1-
943 V3 sequences in the ***V1V3_Raw_TS*** training set were collapsed into one. If identical
944 sequences came from more than one taxon (purple), species-level names of all taxa
945 involved were concatenated (AB). (**e**) The ***V1V3_Supraspecies_TS*** training set includes
946 the same sequences that the ***V1V3_Raw_TS*** training set; however, the headers in the
947 fasta file include the supraspecies taxon (AB) as an extra level between the genus (G)
948 and species taxonomic levels (A, B or AB), as illustrated here.

949 **Figure 3. The *FL_Compilation_TS* training set provides higher classification**
950 **percentages with a lower error rate.** The naïve Bayesian RDP Classifier was used with
951 bootstrap values ranging from 50 to 100. (**a**) The percentage of eHOMD-derived

952 simulated reads classified using the ***FL_eHOMDrefs_TS*** training set (purple) versus the
953 ***FL_Compilation_TS*** training set (orange). (**b**) The percentage of classified reads that
954 were misclassified (i.e., reads for which the assigned taxonomic identity was different
955 than the known identity of the original sequence from which the simulated read was
956 derived).

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

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

974 **ADDITIONAL FILES.**

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

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

979 **Additional File 3**: V1V3_eHOMDSim_250N100.fa. The simulated eHOMD-derived
980 **V1V3_eHOMDSim_250N100** dataset.

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

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

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

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

989 **Additional File 8**: V1V3_sinonasal_SMRT_ASV.fa. Sequences of the resulting 204
990 ASVs in the **FL_sinonasal_SMRT_ASV** dataset trimmed to the V1-V3 region
991 (**V1V3_sinonasal_SMRT_ASV** dataset).

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

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

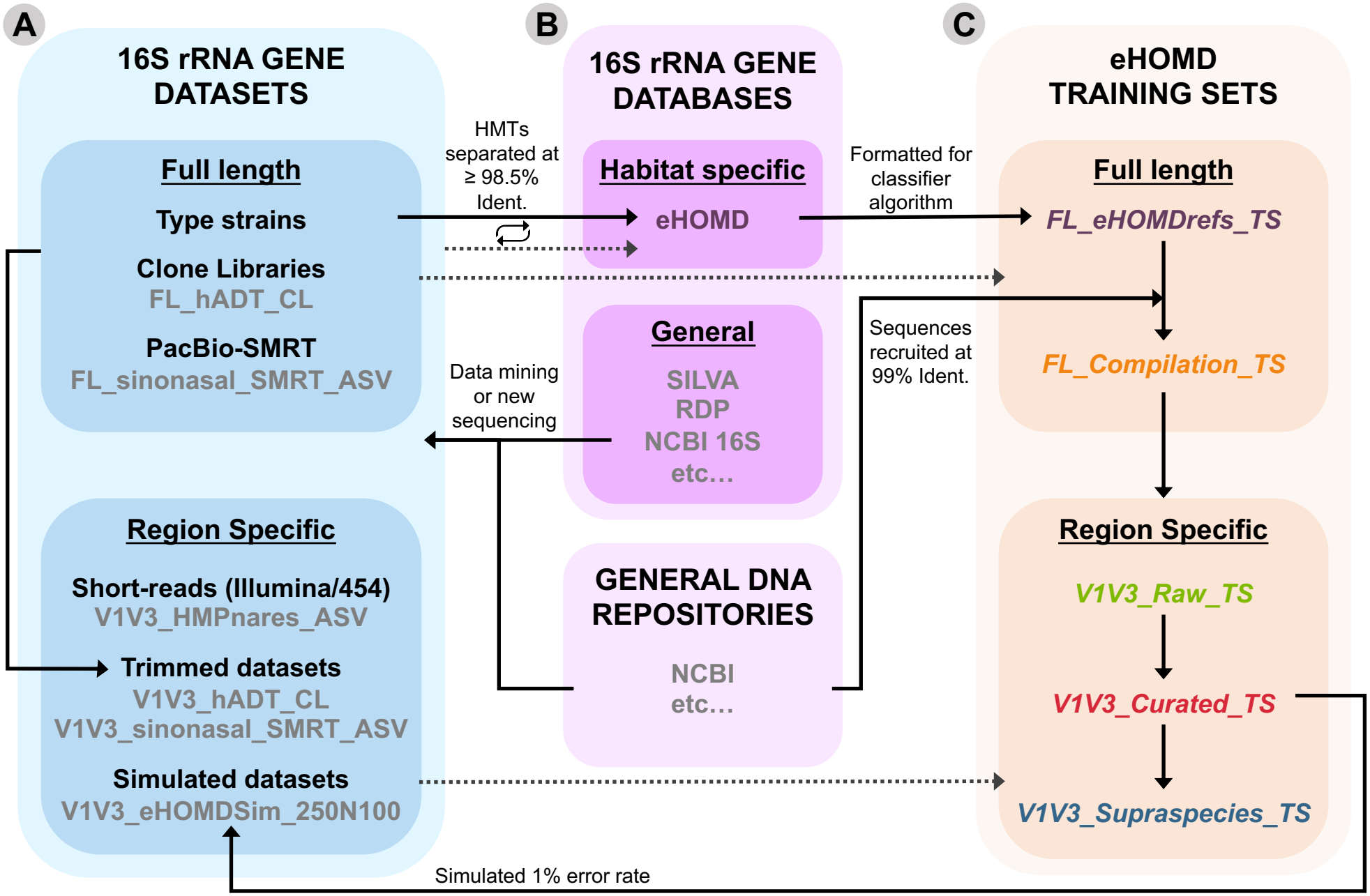
996 **Additional File 11**: retrieveSegment.py. Custom script developed to generate a trimmed
997 version of a training set (in our example to positions 40-880 in the gapped eHOMDrefs
998 alignment) from a compilation of 16S rRNA gene full-length sequences.

999 **Additional File 12**: Simulated_R2_350_R1_20-200.7z. Simulated eHOMD-derived
1000 dataset versions with R2 fixed at 350 bp and R1 fragments from the V3 primer ranging
1001 from 20 to 200 bp.

1002 **Additional File 13**: Simulated_R2_140-350_R1_200.7z. Simulated eHOMD-derived
1003 dataset versions with R1 fixed at 200 bp and R2 fragments from the V1 primer ranging
1004 from 140 to 350 bp.

1005 **Additional File 14**: FL_hADT_CL.fa. Full-length 16S rRNA gene human aerodigestive
1006 tract clone library dataset (**FL_hADT_CL** dataset).

1007 **Additional File 15**: eight2seven.R. Custom-written R function that converts the
1008 dada2::assignTaxonomy() output from eight to seven taxonomic levels for compatibility
1009 with downstream applications.


A
16S rRNA GENE DATASETS
Full length
Type strains
Clone Libraries
FL_hADT_CL
PacBio-SMRT
FL_sinonasal_SMRT_ASV
Region Specific
Short-reads (Illumina/454)
V1V3_HMPnares_ASV
Trimmed datasets
V1V3_hADT_CL
V1V3_sinonasal_SMRT_ASV
Simulated datasets
V1V3_eHOMDSim_250N100
B
16S rRNA GENE DATABASES
Habitat specific
eHOMD
General
SILVA
RDP
NCBI 16S
etc…
GENERAL DNA REPOSITORIES
NCBI
etc…
C
eHOMD TRAINING SETS
Full length
FL_eHOMDrefs_TS
FL_Compilation_TS
Region Specific
V1V3_Raw_TS
V1V3_Curated_TS
V1V3_Supraspecies_TS
HMTs separated at ≥ 98.5% Ident.
Formatted for classifier algorithm
Data mining or new sequencing
Sequences recruited at 99% Ident.
Simulated 1% error rate

A

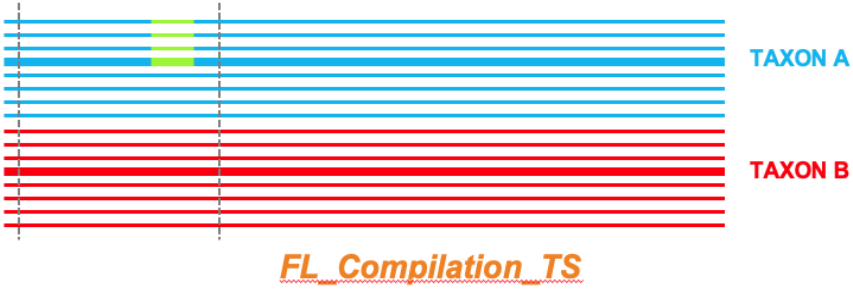
$$P(w_i|T_A) = \frac{1 + P_i}{M + 1}$$

$$P(w_i|T_B) = \frac{0 + P_i}{M + 1}$$

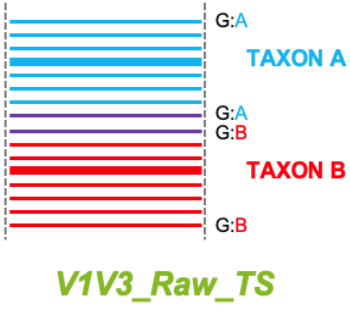
B

$$P(w_i|T_A) = \frac{m(w_i) + P_i}{M + 1}$$

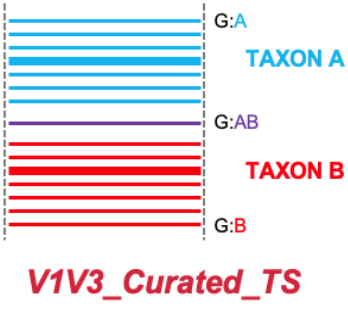
$$P(w_i|T_B) = \frac{m(w_i) + P_i}{M + 1}$$



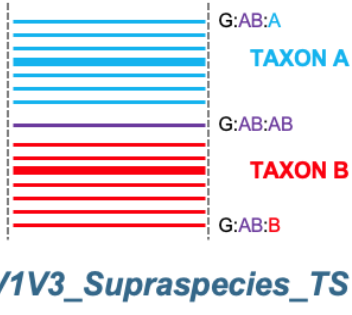
C

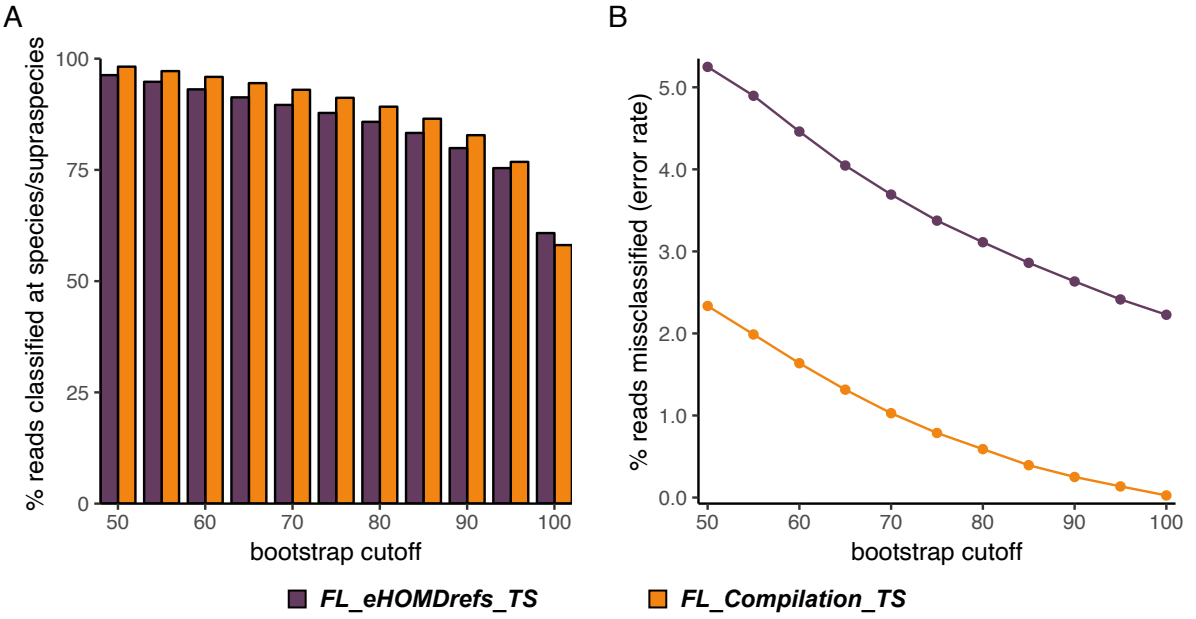


D

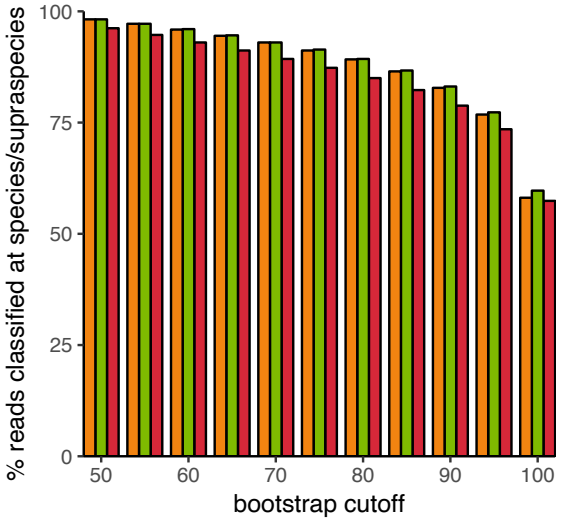


E

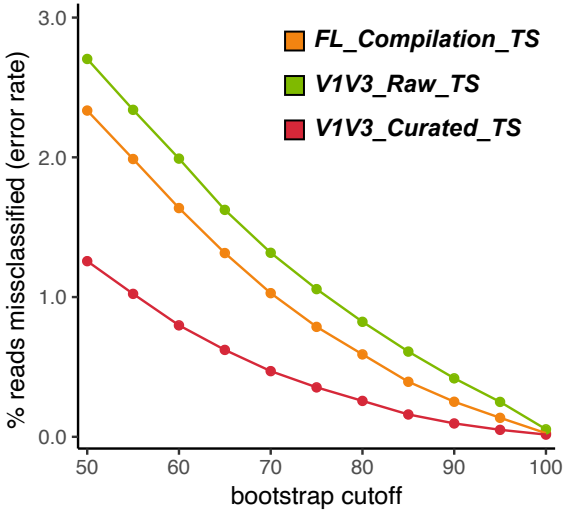



A
% reads classified at species/supraspecies
100
75
50
25
0
50
60
70
80
90
100
bootstrap cutoff
B
% reads missclassified (error rate)
5.0
4.0
3.0
2.0
1.0
0.0
50
60
70
80
90
100
bootstrap cutoff
FL_eHOMDrefs_TS
FL_Compilation_TS

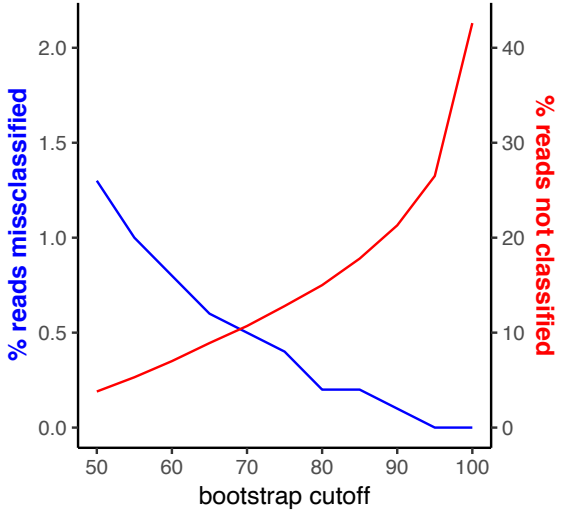
A

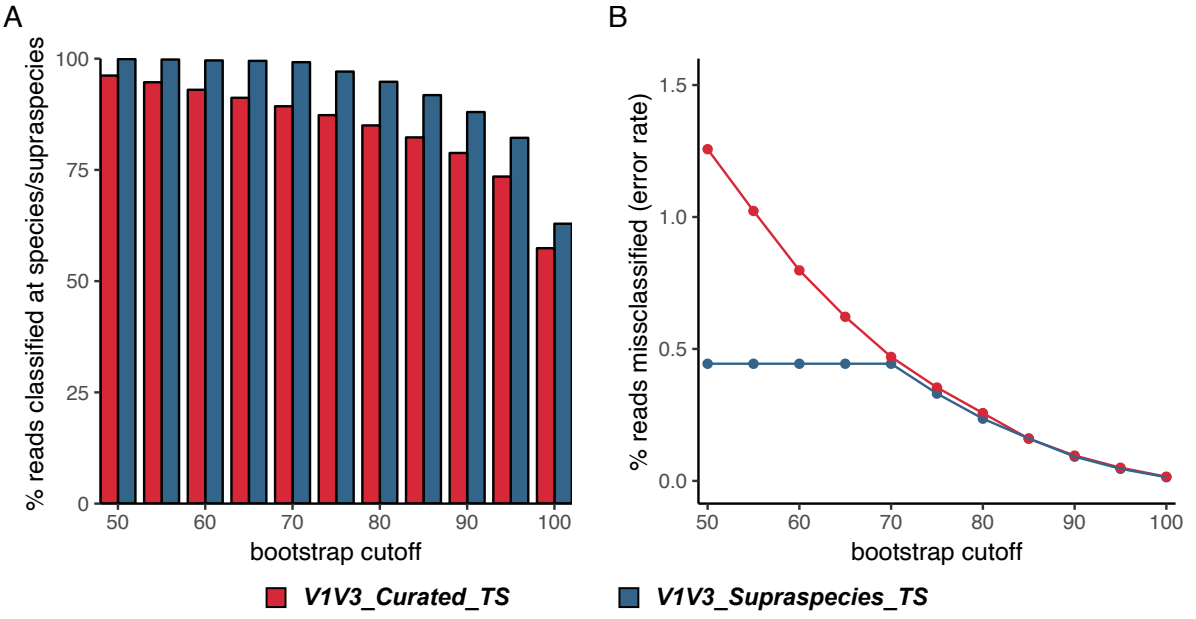

% reads classified at species/supraspecies
100
75
50
25
0
50
60
70
80
90
100
bootstrap cutoff

B


FL_Compilation_TS
V1V3_Raw_TS
V1V3_Curated_TS
% reads missclassified (error rate)
3.0
2.0
1.0
0.0
50
60
70
80
90
100
bootstrap cutoff

C


% reads missclassified
% reads not classified
2.0
1.5
1.0
0.5
0.0
40
30
20
10
0
50
60
70
80
90
100
bootstrap cutoff


A
% reads classified at species/supraspecies
100
75
50
25
0
50
60
70
80
90
100
bootstrap cutoff
B
% reads missclassified (error rate)
1.5
1.0
0.5
0.0
50
60
70
80
90
100
bootstrap cutoff
V1V3_Curated_TS
V1V3_Supraspecies_TS