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

Isabel F. Escapa[a,b,§] (ifescapa@forsyth.org), 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]Department of Molecular Virology and Microbiology, Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, Texas, USA

[d]Department of Pediatrics, Section of Infectious Diseases, Texas Children's Hospital, Houston, Texas, USA

[e]Division of Infectious Diseases, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 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 ecologically coherent entities. Assigning species-level taxonomy to these ASVs is the critical remaining barrier to drawing ecologically/clinically relevant inferences from and comparing data across 16S rRNA gene-based microbiota studies.

Results: To overcome this barrier, we developed a broadly applicable method for constructing a phylogeny-based, high-resolution, habitat-specific training set. When used with the naïve Bayesian Ribosomal Database Project (RDP) Classifier, this training set achieved species/supraspecies-level taxonomic assignment to 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 maintain subgenus resolution. As proof of principle, we developed a V1-V3 region training set for the bacterial microbiota of the human aerodigestive tract using our expanded Human Oral Microbiome Database (eHOMD). In addition, we 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. We also 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. The latter 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; habitat-specific database; microbiota; microbiome; eHOMD; nasal; aerodigestive tract

BACKGROUND.

In microbiota studies of most ecosystems and/or habitats, achieving ecologically and/or clinically relevant results requires species-level identification of constituents. For example, species-level identification is often critically important for host-associated microbial communities because these often include commensal and pathogenic species of the same genus, e.g., [1, 2]. Also, some microbial genera include species that are site specialists and inhabit distinct niches of a given environment [3]. Although high-throughput close-to-full-length 16S rRNA gene sequencing (e.g., circular consensus sequences from PacBio Single Molecule, Real-Time (SMRT) sequencing) and

61 metagenomic whole genome sequencing (WGS) both hold promise for species-level
62 microbiota studies, the current cost per sample limits their utility for studies of more than
63 hundreds of samples. In contrast, the low cost of 16S rRNA gene short-read sequencing
64 makes population-scale (i.e., thousands of samples) molecular epidemiological studies
65 of bacterial microbiota of humans, other animals, plants and the environment broadly
66 feasible. The caveat being that the majority of published short-read 16S rRNA gene
67 sequencing studies use read clustering at a percent similarity that constrains resolution
68 to the genus level, i.e., 97% identity. Indeed, recent reviews on best practices and
69 benchmarking for 16S rRNA gene microbiota studies focus on genus-level operational
70 taxonomic unit (OTU) analysis, e.g., [4-6]. However, newer algorithms, many of which are
71 not based on similarity thresholds, allow single-nucleotide resolution and can parse 16S
72 rRNA gene short-read sequences to species- or strain-level resolution phylotypes, usually
73 called amplicon sequence variants (ASVs) (e.g., MED (Minimal Entropy Decomposition)
74 [7, 8], DADA2 (Divisive Amplicon Denoising Algorithm) [9, 10] and UNOISE [11, 12],
75 among others [13, 14]). Thus, it is vital to develop methods that achieve accurate
76 taxonomic assignment while preserving the sub-genus level resolution of ASVs and
77 permit consistent cross-study comparison.

78 Microbial databases encompassing broad phylogenetic diversity, such as SILVA [15, 16],
79 RDP [17] and Greengenes [18], serve the key role of being applicable to myriad different
80 habitats but this valuable breadth comes with the trade-off of inclusion of taxonomically
81 misannotated 16S rRNA gene sequences. For example, Edgar estimated annotation error
82 rates as high as ~10-17% in these comprehensive databases [19]. (Of note, SILVA and
83 RDP continue to undergo regular updates and contain a broadly expansive and

84 comprehensive record of 16S rRNA gene sequences from all explored habitats, whereas,
85 Greengenes was last updated in 2013.) For habitats that have yet to be deeply
86 interrogated, the access to this breadth outweighs the risk of misclassification due to
87 annotation error. However, once a habitat is sufficiently explored, a habitat-specific
88 database enables accurate fine-level phylogenetic resolution for taxonomic assignment
89 to ASVs [20-30]. Existing habitat-specific databases are constructed with different
90 methods and can be used to assign taxonomy via different approaches. Examples of this
91 include the following: 1) stand-alone habitat-specific databases consisting of curated
92 collections of close-to-full-length 16S rRNA gene sequences compiled both from other
93 repositories and by generating new sequences from the habitat of interest, e.g., eHOMD
94 for the human aerodigestive tract [20, 29], HITdb for the human gut [23] and RIM-DP for
95 rumen [22]; 2) custom addition of compiled sequences from a specific habitat of interest
96 to augment a broad general database, e.g., HBDB for honey bee [21], DictDB for termite
97 and cockroach gut [27], SILVA19Rum for rumen [30] and MiDAS for activated sludge [24,
98 26]; 3) both a general and a habitat-specific database combined in the same pipeline,
99 e.g., a general database followed by a most common ancestors approach with a custom
100 species-level phylogeny of selected human-associated genera with pathogenic members
101 [31] and FreshTrain with the TaxAss workflow for freshwater [28]. Many of these
102 databases are used to train classifiers for taxonomy assignment.

103 The naïve Bayesian RDP Classifier [32] is one of several effective algorithms for
104 assigning taxonomy, all of which require a training set. (For benchmarking and
105 comparison with other methods, see [33-35].) Properly formatted versions of the broad
106 16S rRNA gene databases SILVA, RDP and Greengenes are available to train the most

107 popular implementations of the naïve Bayesian RDP Classifier. The quality of the training
108 set strongly influences taxonomic assignment and habitat-specific training sets have been
109 developed to increase accuracy of taxonomic assignments [21, 27, 33-36]. However, the
110 resolution of available training sets is mostly limited to the genus level. An exception is
111 the manually curated subset of the Greengenes database corresponding to 89 clinically
112 relevant bacterial genera that was used to assign species-level taxonomy of full-length
113 16S rRNA gene sequences of clinical isolates [37]. Notwithstanding, species-level
114 taxonomy assignment of short-read 16S rRNA gene datasets remains a challenge.

115 We hypothesized that starting from a habitat-specific, locally comprehensive, phylogeny-
116 based, high-resolution (separated at ≥ 98.5% identity) set of curated reference sequences
117 with distinct taxonomic names, we could develop a method to rapidly generate a training
118 set to leverage the strength of the naïve Bayesian RDP Classifier to consistently achieve
119 species- or supraspecies (i.e., subgenus)-level taxonomic assignment of ASVs. Here, we
120 describe this method along with improvements to Illumina sequencing strategies to
121 maximize information content of 16S rRNA gene V1-V3 sequencing reads. We show that
122 use of the naïve Bayesian RDP Classifier with a training set in which each taxon is
123 represented by a collection of highly similar sequences that captures the natural variability
124 of each species resulted in accurate species-level taxonomic assignment of short-reads
125 from the V1-V3 region of the 16S rRNA gene. This represents a methodological
126 advancement. Based on our own research interests and availability of the requisite
127 habitat-specific database of reference sequences, we developed and validated short- and
128 long-read training sets for the microbiota of the human aerodigestive tract using our
129 expanded Human Oral Microbiome Database (eHOMD). The eHOMD is a locally

130 comprehensive, high-resolution, phylogeny-based database for the microbiomes of sites
131 within the human aerodigestive tract (mouth, nasal passages, sinuses, throat,
132 esophagus) that also has good coverage of the lower respiratory tract [29]. The eHOMD
133 provisional naming scheme permits cross-study comparison of 16S rRNA gene
134 sequences from both formally named species and as-yet-unnamed or uncultivated
135 species. Using this, we describe a systematic approach for generating training sets that
136 is applicable to any ecosystem/habitat of interest.

137 **RESULTS.**

138 **Identify the 16S rRNA gene regions to optimize species-level identification for taxa**
139 **in the habitat of interest.** The choice of the 16S rRNA gene region(s) for short-read
140 sequencing places an upper bound on the amount of species-level resolution that is
141 achievable within a dataset [38-41]. Therefore, it is key to determine which regions
142 provide the most information for distinguishing bacterial species that are common to that
143 specific ecosystem. For the habitats within the human aerodigestive tract, i.e., the nasal
144 passages, sinuses, throat, mouth, esophagus, as well as the lower respiratory tract, we
145 previously showed that many more taxa are distinguishable at the species level with the
146 V1-V3 region than with the commonly used V3-V4 region of the 16S rRNA gene [29].

147 **A method for achieving highly informative 16S rRNA gene V1-V3 region sequencing**
148 **data using Illumina MiSeq.** The length of the V1-V3 16S rRNA gene region, which
149 corresponds to *E. coli* positions 28 to 517, is highly variable in bacteria. The eHOMD
150 includes 770-plus species level taxa and the eHOMD phylogeny is available at
151 http://www.homd.org/ftp/HOMD_phylogeny/current/. For all but three taxa the V1-V3

152 region size ranged from 431 to 523 base pairs (bp). Three Campylobacter species, with
153 intervening sequences in the V2 region, were up to 653 base pairs in length (**Figure 1A**).
154 This size variability makes it extraordinarily difficult to capture full-length V1-V3
155 sequences for all bacterial taxa using Illumina technology. Even using the 500 or 600
156 cycle MiSeq kits, this is impeded by the decline in sequence quality as the run progresses
157 (**Additional file 1A**). Many analysis pipelines reject read pairs that do not overlap.
158 Excluding taxa with longer V1-V3 regions biases results and generates a false
159 understanding of these microbial communities. To solve these problems, we
160 hypothesized that there are specific subregions within V1-V3 that are most informative
161 such that complete sequencing across the V1-V3 region via overlapping paired reads is
162 unnecessary. To identify the more informative V1-V3 subregions, we evaluated Shannon
163 entropy for each position across all of taxa represented by eHOMD reference sequences
164 (eHOMDrefs) (**Figure 1B**). Based on examination of the entropy values across the V1-
165 V3 region, we postulated that sequencing 300 bp from the V1 primer and 150 bp from the
166 V3 primer would capture the majority of the sequencing diversity needed to maximize
167 species-level taxonomic assignment to the V1-V3 region for species included in eHOMD.

168 Moreover, one advantage of the naïve Bayesian RDP Classifier, which is a k-mer-based
169 Bayesian approach, is that it can provide a single taxonomic assignment when analyzing
170 Illumina pair-end data without requiring overlap of the forward and reverse reads.
171 Therefore, from each primer, V1 and V3, we separately visualized at what read length
172 there was no longer a gain in the percentage of classified sequences using a simulated
173 eHOMD-derived dataset. To accomplish this for primer V3, we fixed the read length from
174 primer V1 at 350 bp (**Figure 1C**) based on the indication in Figure 1B that the first 300 bp

175 captures the majority of sequence variability with the extra 50 bp allowing for variability in
176 region length. Similarly, for primer V1, we fixed the read length from primer V3 at 200 bp
177 (**Figure 1D**). We then determined the percentage of sequences classified at species level
178 at increasing read lengths from the complementary primer using the 998 eHOMDrefs as
179 a training set (***FL_eHOMDrefs_TS***). Based on these assays, species-level assignment
180 plateaued across all bootstrap values at 70 bp from primer V3 (**Figure 1C**). Moreover,
181 species-level assignment started to plateau across all bootstrap values around 210 bp
182 from primer V1 (**Figure 1D**). These results indicated that species-level identification can
183 be achieved with nonoverlapping reads for the 16S rRNA gene V1-V3 region using the
184 naïve Bayesian RDP Classifier, at least for the taxa in eHOMD.

185 Furthermore, these results establish guidance for actual read lengths needed for Illumina
186 sequencing of the V1-V3 16S rRNA gene region. Knowing this, we next determined
187 experimentally what length of high-quality sequences are obtainable using primers V1
188 and V3 with 500 cycles on the Illumina MiSeq. Amplicon-based libraries can be
189 challenging to sequence using Illumina's 2-channel sequencing chemistry since their low
190 diversity hinders correct identification of DNA clusters and accurate base calling [42]. To
191 mitigate this, we changed Read 1 (R1) to start from the V3 primer instead of V1, since
192 clusters are defined very early in an Illumina run (first 4 cycles) [43] and sequences are
193 mostly identical in the first positions immediately 3' of the V1 primer. We also increased
194 the proportion of PhiX from a range of 10-25% up to a range of 30-50%, as recommended
195 in [43]. Together, these steps accomplished a significant improvement in sequence
196 quality (**Additional file 1A-B**).

197 During our experiments, we also observed that variability in library quality, e.g., from low
198 biomass habitats such as the human nostrils, can negatively impact the accuracy of
199 library quantification leading to an inadvertently low percentage of PhiX. Because the
200 informative region from V1 (R2) includes the first 250 bp from the primer (**Figure 1D**), we
201 routinely perform asymmetrical sequencing of R1 and R2 with R1 = 100 bp and R2 = 400
202 bp (e.g., **Additional file 1C**) to avoid having the decrease in sequence quality that is
203 sometimes observed at the end of an Illumina MiSeq run falling within the informative
204 region from the V1 primer (R2) (**Figure 1B**). After trimming poor quality regions from each
205 experimentally-derived read, this generally results in high-quality sequences of 200 to 250
206 bp from primer V1 and 100 bp from primer V3. Thus, for the simulated eHOMD-derived
207 dataset (**V1V3_eHOMDSim_250N100**) created to test each step in the development of
208 the eHOMD training set, we mimicked these Illumina-generated nonoverlapping reads by
209 designing artificial reads consisting in 250 bp from primer V1 and 100 bp from primer V3
210 (**Additional file 2**). (To enhance clarity, the names of datasets appear in **bold** whereas
211 the names of training sets appear in ***bold italics***.)

212 **Compilations of closely-related sequences for each taxon in a training set improve**
213 **the accuracy of species-level taxonomic classification.** Genus-level taxonomic
214 assignment is not an inherent limitation of the naïve Bayesian RDP Classifier. Rather,
215 taxonomic assignment to 16S rRNA gene short reads is limited by both the resolution to
216 which sequences in datasets are parsed and by the nature of the training set used. The
217 former is overcome by using approaches such as oligotyping/MED [7, 8], DADA2 [9, 10]
218 or zero-radius OTU [11, 19] to parse sequence variants at high resolution. We
219 hypothesized that the limitations inherent in training sets could also be overcome.

220 The naïve Bayesian RDP Classifier algorithm indicates that a training set with a larger
221 number of sequences representing each taxon will result in more confident taxonomic
222 assignment [32]. Based on the conditional probability for a member of a taxon (T), the
223 higher the number of possible times a given "k-mer" (word or w_i) exists in the training set,
224 the greater the confidence with which assignment of that taxon is made. Thus, as the
225 number of sequences (M) in the training set increases (**Figure 2A vs. 2B**), the number of
226 accurate assignments should increase. In support of this, Werner and colleagues found
227 more sequence counts result in more assignments [36]. Therefore, to systematically
228 increase M, we used each of the reference sequences in eHOMD, i.e., the eHOMDrefs,
229 as bait to capture closely matching, publicly available sequences and combined the
230 resulting compilation of sequences for each taxon into a close-to-full-length compilation
231 training set (***FL_Compilation_TS***) that reflects the 16S rRNA gene sequence variability
232 for each taxon (**Additional file 3**). We then assessed the percentage of reads in the
233 simulated dataset **V1V3_eHOMDSim_250N100** that classified at the species-level at
234 incremental bootstrap values from 50 to 100 using the training set ***FL_Compilation_TS***
235 (**Figure 3A,** orange bars) compared with the training set ***FL_eHOMDrefs_TS*** (**Figure 3A,**
236 purple bars) and observed an increase with the compilation except at a bootstrap value
237 of 100. We postulate that the additional sequences classified with the
238 ***FL_eHOMDrefs_TS*** TS at a bootstrap of 100 are misclassified sequences. A higher rate
239 of misclassification is expected with the training set ***FL_eHOMDrefs_TS***, since it includes
240 only a few representative sequences for each taxon. Therefore, we next assayed for any
241 misclassifications. At each bootstrap threshold, the percentage of reads that were
242 misclassified (i.e., reads for which the assigned taxonomic identity was different than the

243 known identity of the original sequence from which the simulated read was derived) was
244 at least 50% lower using the training set ***FL_Compilation_TS*** (**Figure 3B**, orange line)
245 than with the training set ***FL_eHOMDrefs_TS*** (**Figure 3B**, purple line). Thus,
246 classification of the dataset **V1V3_eHOMDSim_250N100** showed a reduced error rate
247 and increased confidence level when using a training set consisting of a compilation of
248 closely related sequences rather than one or a few reference sequences for each taxon.

249 **Moving towards an appropriate short-read fragment training set.** The training set
250 ***FL_Compilation_TS*** consisted of close to full-length 16S rRNA gene sequences. We
251 hypothesized that the presence in the training set of k-mers outside the V1-V3 region
252 might lead to misclassifications when using it with a V1-V3 region dataset. Supporting
253 this, when using a training set based on a large general database, trimming reference
254 sequences to match the sequenced region increases the number of sequences that are
255 assigned taxonomy [36]. Therefore, we trimmed the sequences in the training set
256 ***FL_Compilation_TS*** to cover only the V1-V3 region generating training set
257 ***V1V3_Raw_TS*** (**Figure 2C**). Using this there was no gain in the percentage of reads in
258 the simulated dataset **V1V3_eHOMDSim_250N100** classified to the species-level
259 (**Figure 4A**, green bar). Moreover, we observed an increase in the percentage of
260 misclassified reads (**Figure 4B**, green line). Therefore, we next determined why the use
261 of appropriate short-read fragments in this training set paradoxically decreased
262 confidence and accuracy of species-level assignment and resolved the issue.

263 **Combining closely related, indistinguishable taxa into supraspecies decreases the**
264 **error rate for a short-read training set.** Considering the possible explanations for this,

265 we realized that taxa with distinct full-length 16S rRNA gene sequences can have
266 identical V1-V3 sequences. In silico, using only V1-V3, 37 of the ~770 species-level taxa
267 in eHOMD are no longer distinguishable from at least one other species at 100% identity
268 [29]. Therefore, we hypothesized that the observed loss in accuracy using training set
269 ***V1V3 _Raw_TS*** was due to identical sequences with more than one species name, e.g.,
270 *Veillonella parvula* and *Veillonella dispar* [29]. To avoid this problem, we removed
271 duplicate sequences and assigned a combined name, i.e., a supraspecies name, e.g.,
272 *Veillonella* parvula_dispar to the remaining unique sequence (**Figure 2D**). This generated
273 the training set ***V1V3_Curated_TS***, which showed improved accuracy compared to both
274 the ***FL_Compilation_TS*** and ***V1V3_Raw_TS*** (**Figure 4B**, red line). However, this
275 improvement came at a cost of an 0.7 to 4.4 % decrease in the reads assigned
276 supraspecies- or species-level taxonomy at each bootstrap threshold (**Figure 4A**, red
277 bar). This trade-off can be illustrated by graphing at each bootstrap threshold the
278 percentage of reads from the simulated **V1V3_eHOMDSim_250N100** dataset that were
279 misclassified versus confidently assigned species-level taxonomy using the
280 ***V1V3_Curated_TS*** with the naïve Bayesian RDP Classifier (**Figure 4C**).

281 **Maximizing performance of the naïve Bayesian RDP Classifier for subgenus-level**
282 **taxonomic assignment of short-read sequences requires both insertion of**
283 **supraspecies as a taxonomic level and setting a threshold bootstrap**. Two steps
284 were required to reap the benefits of using the supraspecies definition. First, formally
285 inserting supraspecies as a taxonomic level between genus and species in the name
286 header for each sequence in the training set, which yielded the training set
287 ***V1V3_Supraspecies_TS*** (**Figure 2E; Additional File 4**). Second, establishing the

288 bootstrap cut-off at which a sequence is not assigned at the species level so that naïve
289 Bayesian RDP Classifier will then default to supraspecies, rather than defaulting to genus,
290 allowing for a higher percentage of reads to be assigned (**Figure 5A**, blue bar). This latter
291 step preserves subgenus level information encoded in the ASVs. This choice inevitably
292 involves a trade-off between accuracy and percentage of reads classified below genus
293 level, e.g., see **Figure 4C**. For our purposes, we chose a conservative bootstrap value of
294 70 (**Figure 5B,** blue line). With the simulated data, for which the truth is known, this gave
295 an error rate of around 0.05%. Of note, the default bootstrap setting for the naïve
296 Bayesian RDP Classifier is 50 and we would encourage use of a more conservative value
297 for species-level taxonomic assignment.

298 **Using an independent sinonasal dataset validated the performance of the eHOMD**
299 **training set.** We next validated the performance of the eHOMD training set against a
300 dataset of close-to-full-length 16S rRNA gene sequences from sinonasal samples. Very
301 importantly, these sequences are independent of the sequences recruited to build the
302 training set. Earl and colleagues performed PacBio SMRT sequencing of close-to-full-
303 length 16S rRNA genes from multiple sinonasal samples from each of 12 adults without
304 sinus inflammation who were undergoing removal of a pituitary adenoma [44]. Because
305 this dataset was deposited in the SRA, these sequences were not present in the NCBI
306 nr/nt sequence repository (from which sequences were recruited for the eHOMD training
307 set) and therefore not used in the generation of the eHOMD training set. We used DADA2
308 to denoise these sequences and identify the relative abundance of each ASV in this
309 dataset (**FL_sinonasal_SMRT_ASV; Additional File 5**). After which, we followed a

310 three-pronged approach assess the validity of the eHOMD training set for use with this
311 fully independent sinonasal dataset.

312 First, we compared the blastn-generated taxonomic assignment for these ASVs using
313 eHOMD compared to using NCBI 16S Microbial to identify any taxa in the dataset that
314 might be missing in eHOMD, and thus the training set (**Table 1; Additional File 6A**). A
315 detailed comparison (**Additional File 6B**), showed that 50 ASVs, corresponding to 11.1%
316 of the total reads, were differentially assigned by blastn (98.5% identity cutoff) against
317 eHOMDv15.1 versus against NCBI 16S Microbial. Of these 50 ASVs, 33 were assigned
318 to taxa with an HMT (Human Microbial Taxon) # as a provisional species designation.
319 The majority (17/33) were assigned to *Peptoniphilus* sp. HMT-187, which corresponded
320 in the NCBI 16S Microbial blastn output to *Peptoniphilus lacydonensis* and accounted for
321 7.6 % of the total reads. Similarly, *Peptoniphilus* sp. HMT-187 was 11th in total reads
322 across the nostril samples of 210 HMP participants [29]. The new species name
323 *Peptoniphilus lacydonensis* was published in 2018 and is formally recognized [45];
324 therefore, it is now linked to HMT-187 in eHOMD. Of the remaining 17 ASVs, 8 were
325 mismatches between closely related *Streptococcus* species. Only 6 ASVs were
326 nonassigned (NA) using eHOMD, thus, eHOMD included the vast majority (99.7%) of
327 taxa present in the Earl-Mell sinonasal dataset. We note that based on recent additions
328 both eHOMD and NCBI 16S Microbial allowed correct assignment of ASVs to *Lawsonella*
329 *clevelandensis*, which previously might be misclassified by best blastn hit as a Dietzia
330 species, and to *Neisseriaceae* G-1 bacterium HMT-327, which before receiving a stable
331 taxonomic designation in eHOMD might be misclassified by best blastn hit as a
332 *Snodgrassella* species [29].

333 Second, we compared species-level assignment of the dataset
334 **FL_sinonasal_SMRT_ASV** using blastn against eHOMD versus the naïve Bayesian
335 RDP Classifier with the full-length eHOMD training set ***FL_Compilation_TS***. Less than
336 2% (1.91%) of the total reads in the dataset **FL_sinonasal_SMRT_ASV**, which are
337 represented by 25 ASVs, were differentially assigned species-level taxonomy by these
338 approaches (**Table 1; Additional File 6A**). There was no clear pattern to this differential
339 assignment (**Additional File 6C**). This served to validate the use of the full-length
340 eHOMD training set ***FL_Compilation_TS*** and demonstrated that our approach is an
341 effective method for assigning species-level taxonomy to close-to-full-length 16S rRNA
342 gene sequences such as those derived from PacBio SMRT sequencing.

343 Third, having established that our approach is comparable to blastn for species-level
344 assignment to full-length sequences, we used naïve Bayesian RDP Classifier derived
345 assignments with the training set ***FL_Compilation_TS*** as a proxy for the true species
346 representation in the dataset **FL_sinonasal_SMRT_ASV**. We then used these to assess
347 the performance of the training set ***V1V3_Supraspecies_TS*** for assigning species-level
348 taxonomy to the V1-V3 region sequences of the dataset **FL_sinonasal_SMRT_ASV**
349 (**V1V3_sinonasal_SMRT_ASV; Additional File 7**). There were seventeen V1-V3 ASVs
350 that were differentially assigned, corresponding to 1.64% of the reads (**Table 1;**
351 **Additional File 6D**). Of these 17 differentially-assigned ASVs, 5 belonged to a
352 supraspecies, a level not represented in the full-length training set ***FL_Compilation_TS***.
353 Another 4 ASVs, all of which were present at very low relative abundance, had full-length
354 ASVs assigned to *Anaerococcus tetradius*, whereas the V1-V3 region was assigned to
355 *Anaerococcus octavius*, all with bootstrap values ≤ 92 (**Additional File 6A**). In contrast,

356 when 16S rRNA gene sequences were parsed into ASVs based on their full-length, then
357 cut to V1-V3 there was little to no species-level ambiguity (**Additional File 6D**). Thus, the
358 ambiguity observed with V1-V3 16S rRNA short reads most likely occurs at the level of
359 parsing into ASVs, due to fewer informative sequences, rather than at the assignment
360 step with the naïve Bayesian RDP Classifier and eHOMD training set.

361 **Table 1. The eHOMD workflow performed well for species/supraspecies taxonomy**
362 **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

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

365 **The eHOMD training set outperforms both the SILVA and RDP training sets.** Having
366 validated the use of the eHOMD training set with an independent 16S rRNA gene
367 sinonasal dataset (**Table 1**), we next compared the use of our taxonomic assignment
368 workflow with other currently available pipelines that are coupled with the RDP or SILVA
369 databases. We used three different datasets for this: 1) we generated a V1-V3 dataset
370 derived from a collection of human aerodigestive tract 16S rRNA gene clone libraries
371 (**V1V3_hADT_CL; Additional File 8**); 2) the HMP 16S rRNA gene V1-V3 454-sequenced
372 nostril dataset that we previously analyzed [29]; and 3) the close-to-full-length ASVs of
373 the **FL_sinonasal_SMRT_ASV** dataset. We then assigned genus-level taxonomy to all
374 of these using the naïve Bayesian RDP Classifier with a bootstrap threshold of 70 coupled

375 with three different training sets: an eHOMD training set (V1-V3 or FL), RDP16 or
376 SILVA132. The eHOMD training set resulted in a larger percentage of reads assigned to
377 a specific genus for each dataset; however, all three training sets resulted in genus-level
378 assignment of > 90% of the sequences (**Table 2**). In contrast, a striking difference
379 emerged between the different workflows when assigning taxonomy to species level. Our
380 method of coupling the naïve Bayesian RDP Classifier with an eHOMD training set
381 showed superior performance in terms of the percentage of reads classified compared to
382 the exact string match method that is currently implemented in DADA2 in conjunction with
383 SILVA132 or RDP16 (**Table 2**). As would be expected, the exact match algorithm
384 performed much better for close-to-full-length ASVs than for V1-V3 region ASVs. For the
385 V1-V3 region ASVs, it also performed much better with the V1-V3 sequences in the V1-
386 V3 human aerodigestive tract clone library dataset than with the HMP V1-V3 dataset. We
387 speculate this occurred because the close-to-full-length sequences from the human
388 aerodigestive tract clone library (**V1V3_hADT_CL**) dataset are part of both the RDP and
389 SILVA databases, whereas the HMP V1-V3 454 sequences are not.

390 **Table 2. The eHOMD training set is superior for assigning species/supraspecies-**
391 **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*

392 **Table 2 Legend**. *Exact match algorithm; # ASVs derived from the HMP nares V1-V3
393 dataset, as described in [29], constitute the **V1V3_HMPnares_ASV** dataset (**Additional
394 File 9**).

395 A key implication of these data is that our overall method yields comparable species-level
396 results for 16S rRNA gene sequencing of the human aerodigestive tract using V1-V3
397 short-read sequences, which is very cost effective, compared to using close-to-full-length
398 PacBio SMRT sequences. For example, it readily distinguished sequences of
399 *Staphylococcus aureus* and *Staphylococcus epidermidis*. Another implication is that for
400 species-level analysis of the microbiota of habitats lacking a high-resolution, accurate
401 16S rRNA gene database, PacBio SMRT sequencing coupled with the newly available
402 DADA2 PacBio pipeline [46] to generate ASVs followed by blastn against NCBI 16S
403 Microbial can provide effective species-level taxonomic assignment (**Table 1**).

404 **CONCLUSIONS.**

405 Here, we present a systematic approach for generating and validating habitat-specific
406 16S rRNA gene training sets to achieve species/supraspecies-level taxonomic
407 assignment for short- or long-read 16S rRNA gene sequences. We used the naïve
408 Bayesian RDP classifier with our training set; however, such a training set can be used
409 with other taxonomic classifiers [34, 35, 47-54]. Our training set-construction approach
410 includes several methodological advancements. First, we generated clusters of existing
411 close-to-full length 16S rRNA gene sequences at the 99% level (with ≥ 98% coverage)
412 around highly curated reference sequences. This collection of sequences for each taxon
413 represented the currently known sequence variability for each taxon and increased the

414 accuracy of taxonomic assignment using the naïve Bayesian RDP Classifier (**Figure 3**).
415 Second, we added an intermediate taxonomic level between genus and species, which
416 we dubbed supraspecies, to which we assigned taxa in which sequences overlapped
417 between two or more taxonomic clusters. This increased the percentage of sequences
418 assigned subgenus-level taxonomy by preventing the default of consigning difficult-to-
419 assign sequences to the genus level (**Figure 5**). The process of constructing a short-read
420 training set started with the generation of an effective close-to-full-length training set,
421 which could then be used for taxonomic assignment of long-read 16S rRNA gene
422 sequences such as those from clone libraries and from high-throughput long-read
423 sequencing. As a final critical step, we validated the composition of the eHOMD training
424 set using a PacBio SMRT-sequenced sinonasal dataset consisting of sequences
425 independent of those used to build the training set (**Table 1**). Facilitating
426 species/supraspecies-level analysis of 16S rRNA gene sequence datasets paves the way
427 for cost-effective, population-scale molecular epidemiological microbiota studies that can
428 achieve greater ecological/clinical relevance by reaching species level.

429 The successful implementation of a training set for species/supraspecies-level
430 assignment of 16S rRNA gene short-read (e.g., V1-V3) datasets involved a collaborative
431 give-and-take to iteratively optimize the sequencing protocol and the analysis workflow in
432 conjunction with each other. As a result, here, we also provided three specific protocol
433 recommendations for sequencing of the 16S rRNA gene V1-V3 region with the Illumina
434 MiSeq. First, we show that excellent taxonomic assignment can be achieved with
435 nonoverlapping Illumina reads from V1 and V3 using our recommended sequencing
436 protocol and taxonomic assignment workflow. This counters a common misperception

437 that overlapping Illumina reads are required [5, 55]. The naïve Bayesian RDP Classifier
438 tolerates nonoverlapping V1-V3 sequences; however, some other classifiers might not.
439 Second, we demonstrated the value of initiating Read 1 from the V3 reverse primer since
440 R1 usually consists of a longer stretch of high-quality sequence. Third, we provide
441 evidence for a role for increased amounts of PhiX when sequencing from the V1 primer,
442 due to the high-degree of sequence conservation 3' to this primer. This is analogous to
443 the respective recommendations by Illumina and Mitra and colleagues to add PhiX to
444 reach up to or, sometimes, greater than 50% of the total DNA for low diversity samples
445 [43].

446 We note these three limitations to the presented method for constructing a habitat-specific
447 training set. First, capturing the natural sequence variation for the 16S rRNA gene(s) for
448 each taxon is limited by the number of such sequences currently available in public
449 repositories. For example, 140 of the ~770 taxa in eHOMD are represented by four or
450 fewer distinct close-to-full-length sequences in the training set ***FL_Compilation_TS***,
451 whereas, ~630 taxa are represented by five or greater. Second, the comprehensiveness
452 of a training set depends on the comprehensiveness of the phylogeny-based, habitat-
453 specific database from which it is constructed. Therefore, if an ASV is from a species that
454 is absent from the database, then misclassification of that ASV to the mostly closely
455 related species present in the database is possible. This is a known limitation of closed
456 reference-based taxonomic assignment. As such, the known limitations of eHOMD are
457 also limitations of any eHOMD training set [20, 29]. Third, as with databases, training sets
458 also require regular updating over time.

459 Within these limitations, we note several advantages. First, when coupled with a training
460 set built with our method, the k-mer-based naïve Bayesian approach accommodates the
461 natural variability of 16S rRNA gene sequences that exists within many bacterial species
462 enabling high rates of accurate taxonomic assignment. In contrast, this natural variability
463 limits the utility of any exact match algorithm to assigning species-level taxonomy for only
464 those sequences already existing in a training set (**Table 2**). Second, despite all of the
465 known limitations of a single-gene taxonomic indicator, the huge number of 16S rRNA
466 gene sequences from diverse ecosystems available in public repositories supports the
467 utility of the 16S rRNA gene for taxonomic assignment. In contrast, the utility of WGS
468 metagenomic sequencing, which holds the promise of strain-level taxonomic assignment,
469 will remain limited by the number of reads that cannot be mapped until multiple strains
470 genomes are sequenced for each taxon residing in the ecosystem/habitat of interest.

471 Finally, there are several additional recommendations to improve taxonomic assignment
472 achieved with this method. First, we recommend incorporating a stable provisional
473 naming scheme into any habitat-specific database, e.g., the HMTs in eHOMD [20, 29].
474 Second, we recommend validating any training set generated with this method using a
475 habitat-appropriate dataset that is fully independent of the sequences used to construct
476 the training set, as we did with the sinonasal PacBio SMRT-generated dataset. Third, for
477 ASVs that remain unassigned at finer taxonomic levels and are of interest based on their
478 relative abundance and/or prevalence in the population, we recommend two steps. First,
479 query these unassigned ASVs against a broad database such as SILVA, RDP and/or
480 NCBI 16S Microbial via blastn (≥98.5% identity over ≥98% coverage) to ascertain if these
481 belong to a named species. Second, if no such named species match exists, then create

482 a new provisional taxon in your database [20, 29]. This approach will identify candidate
483 new taxa, both named and unnamed, for addition to a habitat-specific database and
484 training set.

485 **METHODS.**

486 **Determination of the 16S rRNA gene V1-V3 sequence length for each taxon in**
487 **eHOMD.** To construct Figure 1A, the 998 eHOMDv15.1 aligned 16S reference sequences
488 (eHOMDrefs) were downloaded in fasta format from eHOMD
489 (http://www.homd.org/?name=seqDownload&type=R) [29]. One eHOMDref was selected
490 for each of the 770 HMTs in eHOMD as a representative of the given HMT. The resulting
491 alignment was trimmed from *E. coli* position 28 to 517 (V1-V3 region) and further trimmed
492 to positions 28-220 (V1-V2 region) and 436-517 (V3 region).

493 **Entropy calculation for the 16S rRNA gene V1-V3 region for each taxon in eHOMD.**
494 To construct Figure 1B, sequences in the eHOMDv15.1 alignment were trimmed from
495 aligned positions 40 to 880. The resulting sequences with alignment gaps are much
496 longer than actual V1-V3 sequences. Shannon entropy for each position in the aligned
497 sequences was estimated with a maximum likelihood method using the "entropy" R
498 package [56]. Because the distribution of ungapped V1-V3 lengths centered around 489
499 bases, a V1-V3 fragment with an ungapped length of 489 (eHOMDref 762_5043) was
500 randomly chosen from the V1-V3 sequence collection as a representative onto which the
501 entropy results were plotted.

502 **Nonoverlapping V1-V3 16S rRNA gene sequencing.** Supragingival pooled plaque
503 samples were deposited into collection tubes containing 150 µL TE buffer. Samples were
504 incubated at 37°C for 3 hrs with 1 µL of Lucigen's Ready-Lyse Lysozyme, and then
505 subsequently processed using Lucigen's MasterPure Complete protocol beginning at
506 step C-Lysis of Fluid or Tissue Samples. The method deviated from the protocol at step
507 4 where the incubation time was increased to 30 mins and steps 5 and 6 were omitted.
508 Samples were precipitated using step F-Precipitation of Total Nucleic Acids and were
509 resuspended in 25 µl of TE buffer instead of the suggested 30 µl. 16S rRNA gene libraries
510 were prepared by targeted amplicon sequencing performed using a custom dual-index
511 protocol [55]. The 16S rRNA gene primers described by Allen and colleagues (**Additional**
512 **File 10)** were used to amplify the V1-V3 region of the 16S rRNA gene [57]. These primers
513 provide broad coverage of 16S rRNA genes while maintaining high sensitivity. The
514 sample libraries were prepared using a 22-cycle PCR reaction to reduce the chimera
515 formation that is common when using a high cycle number. The final PCR products were
516 purified using Ampure XP beads, pooled in equal amounts, and gel purified using the
517 QIAGEN MinElute Gel Extraction Kit. Purified, pooled libraries were quantified using the
518 NEBNext Library Quant Kit for Illumina. Final libraries were sequenced on Illumina®
519 MiSeq™ with a v2 reagent kit (500 cycles) using a 250x250 or 100x400 configuration.
520 The sequencing was performed at a 10 pM loading concentration with the indicated
521 percentage of PhiX added. Quality Phred Score of the reads was assessed using FastQC
522 [58].

523 **Construction of eHOMD-based training datasets for the naïve Bayesian RDP**
524 **Classifier.** Training datasets were constructed in FASTA format as described on the

525 DADA2 official website (https://benjjneb.github.io/dada2/index.html) [9], with taxonomy in
526 the FASTA header line for each individual sequence. Training sets used in this study are
527 described below.

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

531 ***FL_Compilation_TS***. This is the final version of the full-length 16S rRNA gene eHOMD
532 training set, available as **Additional File 3** and for download from eHOMD.org as
533 ***eHOMDv15.1_FL_Compilation_TS.fa.gz***. First, to more precisely calculate the percent
534 identity for recruiting sequences for the training set, we trimmed each of the eHOMDrefs
535 to nucleotides 28 and 1373. This was necessary for two reason: 1) several eHOMDrefs
536 have hanging 5' or 3' ends that, if left in place, would affect the calculation of percent
537 identity and 2) this trimming permitted capture of sequences from the extensive datasets
538 that use the reverse primer at ~1390. Second, each of the 998 trimmed eHOMDrefs was
539 queried against the NCBI nonredundant nucleotide (nr/nt) database using blastn (NCBI
540 BLAST 2.6.0+ package) (https://www.ncbi.nlm.nih.gov/books/NBK279690/) with the
541 parameters -db nr -remote -perc_identity 97. Nucleotide sequences of GenBank IDs with
542 ≥ 97% sequence identity to any of the eHOMDrefs were downloaded in FASTA format
543 using the efetch.fcgi command of NCBI's Entrez Programming Utilities (E-utilities;
544 https://www.ncbi.nlm.nih.gov/books/NBK25501/). The blastn hits were downloaded in the
545 same orientation as the eHOMDrefs in two batches. 1) For subject sequence length
546 between 1000 and 2000 nt, the entire subject sequence was downloaded. 2) For subject

547 sequence length > 2000 bp (e.g., a complete genomic sequence), only the aligned portion
548 of the sequence was downloaded. Sequences < 1000 bp were not downloaded. Third,
549 the 301,794 downloaded sequences, which matched to HMTs at ≥ 97%, were parsed
550 based on their highest sequence percent identity (≥ 99%) and alignment coverage (≥ 98%
551 calculated based on the length of the reference) to any of the eHOMDrefs in a given HMT.
552 The choice of ≥ 99% identity was designed to obtain the centrally conserved set of
553 sequences for each eHOMDref. The choice of ≥ 98% coverage was to ensure that the
554 majority of the close-to-full-length sequence was present. Sequences that matched to
555 multiple HMTs at equal percent identity and coverage were randomly assigned to only
556 one HMT. The ***FL_Compilation_TS*** training set is comprised of a total of 223,144
557 sequences parsed to their corresponding HMT, with a range of 1 to 4004 sequences per
558 HMT.

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

565 ***V1V3_Curated_TS***. Sequences that were identical across V1-V3 in ***V1V3_Raw_TS*** were
566 collapsed into a single sequence with the names of all taxa involved concatenated with a
567 ":" separator. The majority of such concatenations occurred among either the same
568 species, resulting in no name change, or different species of the same genus, resulting

569 in assignment of a concatenated species name. However, there were a number of cases
570 where species from two different genera were involved. These intergenus concatenations
571 were carefully examined case by case. In all but one case, the concatenation was only
572 supported by two to three sequences and, therefore, was deemed unreliable and rejected.
573 After manual examination, only one intergenus concatenation remained. This was
574 between HMT-559 (*Afipia broomeae*) and HMT-597 (*Bradyrhizobium elkanii*) and was
575 supported by 29 sequences. The genus of the concatenated taxa was assigned as
576 *Afipia:Bradyrhizobium* and species as *broomeae:elkanii*.

577 **V1V3_Supraspecies_TS**. This is the final version of the V1-V3 16S rRNA gene eHOMD
578 training set, available as **Additional File 4** and for download from eHOMD.org as
579 eHOMDv15.1_V1V3_Supraspecies_TS.fa.gz
580 (http://www.homd.org/ftp/publication_data/20190709/). The sequences in this training set
581 are the same as those included in the ***V1V3_Curated_TS*** version; only the header
582 information was edited to include the supraspecies notation as a taxonomical level
583 between genus and species. Therefore, instead of the header with seven-levels included
584 in previous versions of the eHOMD training set (i.e.,
585 >Kingdom;Phylum;Class;Order;Family;Genus;Species), the ***V1V3_Supraspecies_TS***
586 training set includes a header with eight levels (i.e.,
587 >Kingdom;Phylum;Class;Order;Family;Genus;Supraspecies;Species). A supraspecies
588 was defined between two or more HMTs when there was a phylogenetic distance less
589 than 0.005 between at least one pair of sequences of the corresponding HMTs. To
590 calculate the distances, V1V3 sequences were aligned with MAFTT v6.935b [59] and
591 treed with FastTree version 2.1.9 [60] with default parameters. Pairwise distances

592 between sequences were calculated as the sum of horizontal branch length between two
593 sequence nodes based on FastTree's default "Jukes-Cantor + CAT" DNA evolution model
594 [60]. The name of the resultant supraspecies is a concatenation of the species names of
595 all HMTs involved, separated by ":" and is assigned at the supraspecies level (seventh
596 level) for all the sequences of the involved HMTs. Sequences that were identical across
597 the V1-V3 region between more than one taxon were assigned the concatenated name
598 both at the supraspecies and species levels (e.g.,
599 >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Coryne-
600 bacterium;accolens:macginleyi:tuberculostearicum;accolens:macginleyi:tuberculostearic
601 um). Whereas, the rest of the reads for those taxa were named at the supraspecies level
602 as the concatenated name and maintained their unique identifier at the species level (e.g.,
603 >Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Coryne-
604 bacterium;accolens:macginleyi:tuberculostearicum;accolens). For taxa not requiring
605 concatenations (see the ***V1V3_Curated_TS*** description), the species designation is
606 repeated at the supraspecies level such that the seventh- and eighth-level designations
607 are identical (e.g., >Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Dolo-
608 sigranulum;pigrum;pigrum).

609 **Generation of a V1-V3 16S rRNA gene simulated dataset derived from eHOMD.**
610 Simulated reads were generated from each unique sequence in the ***V1V3_Curated_TS***
611 training set. During read generation, one percent of the bases in each unique initial
612 sequence were randomly selected and changed into a different base to simulate a 1%
613 error rate. The resulting dataset contains 19,480 simulated sequences, each of them
614 labelled with the ID of the parent eHOMDref sequence from which they were derived.

615 Then, each error-simulated read was trimmed into two fragments to simulate Illumina pair-
616 end reads (R1 starting from the V3 primer and R2 starting from the V1 primer). We
617 generated multiple length configurations of both the R1 and R2 fragments. First, fixing R2
618 at 350 bp, we generated R1 fragments from the V3 primer ranging from 20 to 200 bp
619 (Simulated_R2_350_R1_20-200, **Additional File 12**). Second, fixing R1 at 200 bp, we
620 generated R2 fragments from the V1 primer ranging from 140 to 350 bp
621 (Simulated_R2_140-350_R1_200, **Additional File 13**). Prior to analysis with the naïve
622 Bayesian RDP Classifier, paired reads were connected with a "10N" linker to form one
623 fragment ordered R2-10N-R1 (**Figure 1C** and **1D**). Subsequent evaluation of the steps
624 for generating the eHOMD training sets in Figures 3 to 5 was done using the configuration
625 R2(250 bp)-10N-R1(100 bp) **(V1V3_eHOMDSim_250N100; Additional File 2**).

626 **Reanalysis of a sinonasal PacBio-SMRT-sequenced full-length 16S rRNA gene**
627 **dataset.** Earl and colleagues kindly provided the circular consensus sequences (CCS)
628 already demultiplexed, labeled by sample, pooled together and, then, filtered based on
629 length distribution, terminal matches to the primer sequences, not aligning to a provided
630 host or background genome sequence and with a cumulative expected error (EE < 1)
631 [44]. We then used the DADA2 PacBio pipeline to denoise these CCS and identify the
632 relative abundance of each ASV in the dataset; i.e., we used the learnErrors() function
633 with errorEstimationFunction=PacBioErrfun and BAND_SIZE=32 [9, 46]. The sequences
634 of the resulting 204 ASVs are in the **FL_sinonasal_SMRT_ASV** dataset (**Additional File**
635 **5**). The V1-V3 region was trimmed using bedtools getfasta with default parameters
636 (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/) and reads were
637 further trimmed into two fragments to simulate Illumina pair-end reads using the

638 configuration R2(250 bp)-10N-R1(100 bp), as described above
639 (**V1V3_sinonasal_SMRT_ASV**; **Additional File 7**).

640 **Generating a test V1-V3 human aerodigestive tract (hADT) microbiota dataset from**
641 **full-length 16S rRNA gene clone libraries (CL).** We previously used the close-to-full-
642 length 16S rRNA gene sequences from clone library-based microbiota studies of the
643 human aerodigestive tract, as described in Supplemental Text S1 of [29]: Segre-Kong
644 nostril (SKn) [61-66], Pei-Blaser [67, 68], Harris-Pace [69], van der Gast-Bruce [70],
645 Flanagan-Bristow [71] and Perkins-Angenent [72]. Here, we compiled these into one
646 dataset along with clones from NCBI PopSet UIDs 399192397, 399202217, 399199823,
647 399197584, 399194446, 399189902, 399186216, 399183739, 399182414, 399179617,
648 399175646 and 399173254 [73]. Aligned eHOMDrefs (eHOMDv15.1) sequences were
649 trimmed from *E. coli.* position 28-1373 and used to query this compiled dataset via blastn.
650 We retained 27,816 sequences that hit with 100% coverage and ≥ 99.5% identity to 401
651 HMTs as the full-length human aerodigestive tract clone library dataset (**FL_hADT_CL;**
652 **Additional File 14**). Of these, 5254 (18.9%) matched to more than one HMT, whereas
653 22,562 (81.1%) unambiguously matched to single HMTs. Sequences in this full-length CL
654 dataset were then aligned using MAFTT v6.935b with default parameters [59]. Segments
655 corresponding to the V1-V3 region were extracted based on positions of V1-V3 in the
656 alignment (V1V3_hADT_CL, **Additional File 8**) using bedtools getfasta with default
657 parameters (bedtools version 2.26.0, https://bedtools.readthedocs.io/en/latest/).

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

660 dada2::assignTaxonomy() function in R (an implementation of the algorithm with a kmer
661 size of 8 and 100 bootstrap iterations) with the specified training set and with
662 outputBootstraps=TRUE and the indicated minBoot value [9, 32]. In addition to our
663 eHOMD training sets, we used the RDP16 (***rdp_train_set_16.fa.gz***) and SILVA132
664 (***silva_nr_v132_train_set.fa.gz***) genus-level training sets
665 (https://benjjneb.github.io/dada2/training.html). Since the eHOMD training set
666 ***V1V3_Supraspecies_TS*** generates an output with eight taxonomic levels that might not
667 be compatible with common downstream applications that accept only seven levels, we
668 provide a custom-written R function (**Additional File 15**) that converts the output from
669 eight to seven levels as follows. First, if a sequence is not assigned at species level (i.e.,
670 the eighth level) at a given threshold (output as "NA") then the eighth level is replaced
671 with the classification result and bootstrap value from the corresponding seventh level.
672 Next, the seventh level is deleted and only the resultant merged eighth level that can
673 contain either species or supraspecies information is reported. Please note that this
674 merged last supraspecies-or-species level is labeled simply as species.

675 **Taxonomic assignment with the DADA2 exact match.** ASVs and CL sequences were
676 assigned species-level taxonomy with the RDP16 (***rdp_species_assignment_16.fa.gz***)
677 and SILVA132 (***silva_species_assignment_v132.fa.gz***) training set files downloaded
678 from https://benjjneb.github.io/dada2/training.html using the dada2::assignSpecies()
679 function in R with allowMultiple=TRUE [9].

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

682 command used with max_target_seqs 1, hits with < 98.5% identity were considered
683 nonassigned. ASVs (**FL_sinonasal_SMRT_ASV**) derived from the sinonasal PacBio
684 SMRT-sequenced dataset were also assigned taxonomy using blastn against two
685 databases: the NCBI 16S Microbial database was downloaded from
686 ftp://ftp.ncbi.nlm.nih.gov/blast/db/ on January 2019 [74] and eHOMDv15.1 was converted
687 to a BLAST database using "makeblastdb" from the NCBI BLAST 2.6.0+ package.

688 **DECLARATIONS.**

689 **Ethics approval and consent to participate.** All participants provided informed consent
690 and samples used to generate the V1-V3 region 16S rRNA dataset in Additional File 1
691 were collected under a protocol (#13-14 to FED) approved by the Forsyth Institutional
692 Review Board**.**

693 **Consent for publication.** Not applicable.

694 **Availability of data and material.** All data are included as additional files. The following
695 are also available for download at eHOMD.org
696 (http://www.homd.org/ftp/publication_data/20190709/): 1) the simulated eHOMD derived
697 dataset (**Additional File 2** as V1V3_eHOMDSim_250N100.fa), 2) both eHOMD training
698 set files (**Additional File 3** as eHOMDv15.1_FL_Compilation_TS.fa.gz; and **Additional**
699 **File 4** as eHOMTv15.1_V1V3_Supraspecies_TS.fa.gz) and 3) both custom scripts
700 (**Additional File 11** as retrieveSegment.py and **Additional File 15** as eight2seven.R).

701 **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.

723 **REFERENCES**

724 1. Egerton S, Culloty S, Whooley J, Stanton C, Ross RP: **The Gut Microbiota of**
725 **Marine Fish.** *Front Microbiol* 2018, **9:**873.
726 2. Berendsen RL, van Verk MC, Stringlis IA, Zamioudis C, Tommassen J, Pieterse
727 CM, Bakker PA: **Unearthing the genomes of plant-beneficial Pseudomonas**
728 **model strains WCS358, WCS374 and WCS417.** *BMC Genomics* 2015, **16:**539.
729 3. Welch JLM, Dewhirst FE, Borisy GG: **Biogeography of the Oral Microbiome:**
730 **The Site-Specialist Hypothesis.** *Annu Rev Microbiol* 2019.
731 4. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M,
732 Podar M, Quince C, Hall N: **A comprehensive benchmarking study of**
733 **protocols and sequencing platforms for 16S rRNA community profiling.** *BMC*
734 *Genomics* 2016, **17:**55.
735 5. Pollock J, Glendinning L, Wisedchanwet T, Watson M: **The Madness of**
736 **Microbiome: Attempting To Find Consensus "Best Practice" for 16S**
737 **Microbiome Studies.** *Appl Environ Microbiol* 2018, **84**.
738 6. Almeida A, Mitchell AL, Tarkowska A, Finn RD: **Benchmarking taxonomic**
739 **assignments based on 16S rRNA gene profiling of the microbiota from**
740 **commonly sampled environments.** *Gigascience* 2018, **7**.
741 7. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML:
742 **Oligotyping: Differentiating between closely related microbial taxa using 16S**
743 **rRNA gene data.** *Methods in ecology and evolution / British Ecological Society*
744 2013, **4**.

745 8. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML: **Minimum**
746 **entropy decomposition: unsupervised oligotyping for sensitive partitioning**
747 **of high-throughput marker gene sequences.** *The ISME journal* 2015, **9:**968-
748 979.
749 9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP: **DADA2:**
750 **High-resolution sample inference from Illumina amplicon data.** *Nat Methods*
751 2016, **13:**581-583.
752 10. Callahan BJ, McMurdie PJ, Holmes SP: **Exact sequence variants should**
753 **replace operational taxonomic units in marker-gene data analysis.** *ISME J*
754 2017, **11:**2639-2643.
755 11. Edgar RC: **UNOISE2: improved error-correction for Illumina 16S and ITS**
756 **amplicon sequencing.** *bioRxiv* 2016**:**081257.
757 12. Edgar RC: **Updating the 97% identity threshold for 16S ribosomal RNA OTUs.**
758 *Bioinformatics* 2018, **34:**2371-2375.
759 13. Tikhonov M, Leach RW, Wingreen NS: **Interpreting 16S metagenomic data**
760 **without clustering to achieve sub-OTU resolution.** *ISME J* 2015, **9:**68-80.
761 14. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z,
762 Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R: **Deblur Rapidly**
763 **Resolves Single-Nucleotide Community Sequence Patterns.** *mSystems* 2017,
764 **2**.
765 15. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner
766 FO: **The SILVA ribosomal RNA gene database project: improved data**
767 **processing and web-based tools.** *Nucleic Acids Res* 2013, **41:**D590-596.

768 16. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies
769 J, Ludwig W, Glockner FO: **The SILVA and "All-species Living Tree Project**
770 **(LTP)" taxonomic frameworks.** *Nucleic Acids Res* 2014, **42:**D643-648.
771 17. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro
772 A, Kuske CR, Tiedje JM: **Ribosomal Database Project: data and tools for high**
773 **throughput rRNA analysis.** *Nucleic Acids Res* 2014, **42:**D633-642.
774 18. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A,
775 Andersen GL, Knight R, Hugenholtz P: **An improved Greengenes taxonomy**
776 **with explicit ranks for ecological and evolutionary analyses of bacteria and**
777 **archaea.** *ISME J* 2012, **6:**610-618.
778 19. Edgar R: **Taxonomy annotation and guide tree errors in 16S rRNA databases.**
779 *PeerJ* 2018, **6:**e5030.
780 20. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A,
781 Wade WG: **The human oral microbiome.** *J Bacteriol* 2010, **192:**5002-5017.
782 21. Newton IL, Roeselers G: **The effect of training set on the classification of**
783 **honey bee gut microbiota using the Naive Bayesian Classifier.** *BMC Microbiol*
784 2012, **12:**221.
785 22. Seedorf H, Kittelmann S, Henderson G, Janssen PH: **RIM-DB: a taxonomic**
786 **framework for community structure analysis of methanogenic archaea from**
787 **the rumen and other intestinal environments.** *PeerJ* 2014, **2:**e494.
788 23. Ritari J, Salojarvi J, Lahti L, de Vos WM: **Improved taxonomic assignment of**
789 **human intestinal 16S rRNA sequences by a dedicated reference database.**
790 *BMC Genomics* 2015, **16:**1056.

791 24. McIlroy SJ, Saunders AM, Albertsen M, Nierychlo M, McIlroy B, Hansen AA, Karst
792 SM, Nielsen JL, Nielsen PH: **MiDAS: the field guide to the microbes of**
793 **activated sludge.** *Database (Oxford)* 2015, **2015:**bav062.
794 25. Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR,
795 Cole JR, Midgley DJ, Tran-Dinh N: **Fungal identification using a Bayesian**
796 **classifier and the Warcup training set of internal transcribed spacer**
797 **sequences.** *Mycologia* 2016, **108:**1-5.
798 26. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM,
799 Albertsen M, Nielsen PH: **MiDAS 2.0: an ecosystem-specific taxonomy and**
800 **online database for the organisms of wastewater treatment systems**
801 **expanded for anaerobic digester groups.** *Database (Oxford)* 2017, **2017**.
802 27. Mikaelyan A, Kohler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A:
803 **Classifying the bacterial gut microbiota of termites and cockroaches: A**
804 **curated phylogenetic reference database (DictDb).** *Syst Appl Microbiol* 2015,
805 **38:**472-482.
806 28. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD: **TaxAss: Leveraging a**
807 **Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution.**
808 *mSphere* 2018, **3**.
809 29. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP: **New Insights**
810 **into Human Nostril Microbiome from the Expanded Human Oral Microbiome**
811 **Database (eHOMD): a Resource for the Microbiome of the Human**
812 **Aerodigestive Tract.** *mSystems* 2018, **3**.

813 30. Henderson G, Yilmaz P, Kumar S, Forster RJ, Kelly WJ, Leahy SC, Guan LL,
814 Janssen PH: **Improved taxonomic assignment of rumen bacterial 16S rRNA**
815 **sequences using a revised SILVA taxonomic framework.** *PeerJ* 2019,
816 **7:**e6496.
817 31. Conlan S, Kong HH, Segre JA: **Species-level analysis of DNA sequence data**
818 **from the NIH Human Microbiome Project.** *PLoS One* 2012, **7:**e47075.
819 32. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid**
820 **assignment of rRNA sequences into the new bacterial taxonomy.** *Appl*
821 *Environ Microbiol* 2007, **73:**5261-5267.
822 33. Lan Y, Wang Q, Cole JR, Rosen GL: **Using the RDP classifier to predict**
823 **taxonomic novelty and reduce the search space for finding novel organisms.**
824 *PLoS One* 2012, **7:**e32491.
825 34. Vinje H, Liland KH, Almoy T, Snipen L: **Comparing K-mer based methods for**
826 **improved classification of 16S sequences.** *BMC Bioinformatics* 2015, **16:**205.
827 35. Murali A, Bhargava A, Wright ES: **IDTAXA: a novel approach for accurate**
828 **taxonomic classification of microbiome sequences.** *Microbiome* 2018, **6:**140.
829 36. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG,
830 Angenent LT, Knight R, Ley RE: **Impact of training sets on classification of**
831 **high-throughput bacterial 16s rRNA gene surveys.** *ISME J* 2012, **6:**94-103.
832 37. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S,
833 Nadarajan R, Brodie EL, Lynch SV: **Use of 16S rRNA gene for identification of**
834 **a broad range of clinically relevant bacterial pathogens.** *PLoS One* 2015,
835 **10:**e0117617.

836 38. Kumar PS, Brooker MR, Dowd SE, Camerlengo T: **Target region selection is a**
837 **critical determinant of community fingerprints generated by 16S**
838 **pyrosequencing.** *PLoS One* 2011, **6:**e20956.
839 39. Yang B, Wang Y, Qian P-Y: **Sensitivity and correlation of hypervariable**
840 **regions in 16S rRNA genes in phylogenetic analysis.** *BMC bioinformatics* 2016,
841 **17:**135-135.
842 40. Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, Zhang H, Xiong Z, Xue Y, Tu J, Lu
843 Z: **Evaluation of different 16S rRNA gene V regions for exploring bacterial**
844 **diversity in a eutrophic freshwater lake.** *Sci Total Environ* 2018, **618:**1254-
845 1267.
846 41. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI:
847 **The effect of 16S rRNA region choice on bacterial community metabarcoding**
848 **results.** *Scientific Data* 2019, **6:**190007.
849 42. **Best practices for low diversity sequencing on the NextSeq and MiniSeq**
850 **systems** [https://support.illumina.com/bulletins/2016/03/best-practices-for-
851 lowdiversity-sequencing-on-the-nextseq-and-miniseq-systems.html]
852 43. 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 44. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN,
856 Workman AD, Blasetti M, Sen B, et al: **Species-level bacterial community**
857 **profiling of the healthy sinonasal microbiome using Pacific Biosciences**
858 **sequencing of full-length 16S rRNA genes.** *Microbiome* 2018, **6:**190.

859 45. Beye M, Bakour S, Le Dault E, Rathored J, Michelle C, Cadoret F, Raoult D,
860 Fournier PE: **Peptoniphilus lacydonensis sp. nov., a new human-associated**
861 **species isolated from a patient with chronic refractory sinusitis.** *New*
862 *Microbes New Infect* 2018, **23:**61-69.
863 46. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK,
864 Dougherty MK: **High-throughput amplicon sequencing of the full-length 16S**
865 **rRNA gene with single-nucleotide resolution.** *Nucleic Acids Res* 2019.
866 47. Nguyen NP, Mirarab S, Liu B, Pop M, Warnow T: **TIPP: taxonomic identification**
867 **and phylogenetic profiling.** *Bioinformatics* 2014, **30:**3548-3555.
868 48. Allard G, Ryan FJ, Jeffery IB, Claesson MJ: **SPINGO: a rapid species-classifier**
869 **for microbial amplicon sequences.** *BMC Bioinformatics* 2015, **16:**324.
870 49. Edgar RC: **SINTAX: a simple non-Bayesian taxonomy classifier for 16S and**
871 **ITS sequences.** *bioRxiv* 2016**:**074161.
872 50. Gao X, Lin H, Revanna K, Dong Q: **A Bayesian taxonomic classification**
873 **method for 16S rRNA gene sequences with improved species-level accuracy.**
874 *BMC Bioinformatics* 2017, **18:**247.
875 51. Liland KH, Vinje H, Snipen L: **microclass: an R-package for 16S taxonomy**
876 **classification.** *BMC Bioinformatics* 2017, **18:**172.
877 52. Matias Rodrigues JF, Schmidt TSB, Tackmann J, von Mering C: **MAPseq: highly**
878 **efficient k-mer search with confidence estimates, for rRNA sequence**
879 **analysis.** *Bioinformatics* 2017, **33:**3808-3810.
880 53. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA,
881 Gregory Caporaso J: **Optimizing taxonomic classification of marker-gene**

882 **amplicon sequences with QIIME 2's q2-feature-classifier plugin.** *Microbiome*
883 2018, **6:**90.
884 54. Zheng Q, Bartow-McKenney C, Meisel JS, Grice EA: **HmmUFOtu: An HMM and**
885 **phylogenetic placement based ultra-fast taxonomic assignment and OTU**
886 **picking tool for microbiome amplicon sequencing studies.** *Genome Biol* 2018,
887 **19:**82.
888 55. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD: **Development of**
889 **a dual-index sequencing strategy and curation pipeline for analyzing**
890 **amplicon sequence data on the MiSeq Illumina sequencing platform.** *Appl*
891 *Environ Microbiol* 2013, **79:**5112-5120.
892 56. Hausser J, Strimmer K: **Entropy inference and the James-Stein estimator, with**
893 **application to nonlinear gene association networks.** *J Mach Learn Res*
894 2009**:**1469-1484.
895 57. Allen HK, Bayles DO, Looft T, Trachsel J, Bass BE, Alt DP, Bearson SM, Nicholson
896 T, Casey TA: **Pipeline for amplifying and analyzing amplicons of the V1-V3**
897 **region of the 16S rRNA gene.** *BMC Res Notes* 2016, **9:**380.
898 58. Wingett SW, Andrews S: **FastQ Screen: A tool for multi-genome mapping and**
899 **quality control.** *F1000Res* 2018, **7:**1338.
900 59. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid**
901 **multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids*
902 *Res* 2002, **30:**3059-3066.

903 60. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution**
904 **trees with profiles instead of a distance matrix.** *Mol Biol Evol* 2009, **26:**1641-
905 1650.
906 61. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG,
907 Blakesley RW, Murray PR, Green ED, et al: **Topographical and temporal**
908 **diversity of the human skin microbiome.** *Science* 2009, **324:**1190-1192.
909 62. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley
910 EC, Komarow HD, Murray PR, et al: **Temporal shifts in the skin microbiome**
911 **associated with disease flares and treatment in children with atopic**
912 **dermatitis.** *Genome research* 2012, **22:**850-859.
913 63. Oh J, Conlan S, Polley EC, Segre JA, Kong HH: **Shifts in human skin and nares**
914 **microbiota of healthy children and adults.** *Genome medicine* 2012, **4:**77.
915 64. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos
916 E, Park M, Kong HH, Segre JA: **Topographic diversity of fungal and bacterial**
917 **communities in human skin.** *Nature* 2013, **498:**367-370.
918 65. Oh J, Freeman AF, Park M, Sokolic R, Candotti F, Holland SM, Segre JA, Kong
919 HH: **The altered landscape of the human skin microbiome in patients with**
920 **primary immunodeficiencies.** *Genome research* 2013, **23:**2103-2114.
921 66. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA: **Biogeography and**
922 **individuality shape function in the human skin metagenome.** *Nature* 2014,
923 **514:**59-64.
924 67. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ: **Bacterial biota in the**
925 **human distal esophagus.** *Proc Natl Acad Sci U S A* 2004, **101:**4250-4255.

926 68. Pei Z, Yang L, Peek RM, Jr Levine SM, Pride DT, Blaser MJ: **Bacterial biota in**
927 **reflux esophagitis and Barrett's esophagus.** *World J Gastroenterol* 2005,
928 **11:**7277-7283.
929 69. Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess
930 H, Deterding RR, Accurso FJ, Pace NR: **Molecular identification of bacteria in**
931 **bronchoalveolar lavage fluid from children with cystic fibrosis.** *Proc Natl*
932 *Acad Sci U S A* 2007, **104:**20529-20533.
933 70. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW,
934 Carroll MP, Parkhill J, Bruce KD: **Partitioning core and satellite taxa from within**
935 **cystic fibrosis lung bacterial communities.** *ISME J* 2011, **5:**780-791.
936 71. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P,
937 DeSantis TZ, Andersen GL, Wiener-Kronish JP, Bristow J: **Loss of bacterial**
938 **diversity during antibiotic treatment of intubated patients colonized with**
939 **Pseudomonas aeruginosa.** *J Clin Microbiol* 2007, **45:**1954-1962.
940 72. Perkins SD, Woeltje KF, Angenent LT: **Endotracheal tube biofilm inoculation of**
941 **oral flora and subsequent colonization of opportunistic pathogens.** *Int J Med*
942 *Microbiol* 2010, **300:**503-511.
943 73. Shelef KM: **Ecology in the dentist's chair: patterns of biogeography and**
944 **stability in human subgingival microbial communities.** Stanford University,
945 Department of Biology; 2013.
946 74. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,
947 Robbertse B, Smith-White B, Ako-Adjei D, et al: **Reference sequence (RefSeq)**

948 **database at NCBI: current status, taxonomic expansion, and functional**
949 **annotation.** *Nucleic Acids Res* 2016, **44:**D733-745.

950

951

952

953 **FIGURE LEGENDS.**

954 **Figure 1. 16S rRNA gene V1-V3 region sequences do not require overlap to provide**
955 **maximal information for human aerodigestive tract-associated bacteria.** (**a**) Rank
956 order of eHOMD taxa based on the nucleotide length of regions V1-V3 (purple), V1-V2
957 (orange) and V3 of the 16S rRNA gene. (**b**) Shannon Entropy (H) across the 16S rRNA
958 gene V1-V3 region for all taxa in eHOMD. For easier visualization, bars are color-coded
959 in gray scale based on their entropy values, i.e., the taller a bar is the darker it is. The
960 percentage of eHOMD-derived simulated reads that can be classified with the
961 ***FL_eHOMDrefs*** training set at bootstrap values from 70 to 100 (see key) using (**c**) a fixed
962 read length of 350 bp from primer V1 and variable read lengths from primer V3 or (**d**) a
963 fixed read length of 200 bp from primer V3 and variable read lengths from primer V1.

964 **Figure 2. Schematic representation of the steps to generate sequential habitat-**
965 **specific training sets.** (**a**) The ***FL_eHOMDrefs_TS*** training set contains all full-length
966 eHOMDrefs (thick lines) from eHOMDv15.1 together with their respective taxonomic
967 assignment. When only one read represents each taxon (M=1) a given k-mer (green
968 fragment) can only be either present (1) or absent (0). (**b**) A higher number of sequences
969 per taxon (M) allows for better resolution on the assignment, with the presence of a given
970 k-mer (green fragment) across each cluster of reads (w_i) being represented as a
971 proportion out of the total number of reads in that taxon (M). Therefore, to better represent
972 the known sequence diversity of the 16S rRNA gene(s) for each taxon, the training set
973 ***FL_Compilation_TS*** includes clusters of sequences (thin lines) recovered from the NCBI
974 nonredundant nucleotide (nr/nt) database that matched with 99% identity and ≥ 98%

975 coverage (see methods) to each eHOMDref (thick line). (**c**) The training set
976 ***V1V3_Raw_TS*** is a V1-V3 trimmed version of the ***FL_Compilation_TS*** training set. The
977 schematic illustrates how trimming to this region leads to identical reads (purple lines)
978 having two different taxonomic designations. Here, G is genus and species are labeled
979 as A or B. (**d**) To construct the ***V1V3_Curated_TS*** training set, identical V1-V3 sequences
980 in the ***V1V3_Raw_TS*** training set were collapsed into one. If identical sequences came
981 from more than one taxon (purple), species-level names of all taxa involved were
982 concatenated (AB). (**e**) The ***V1V3_Supraspecies_TS*** training set includes the same
983 sequences that the ***V1V3_Raw_TS*** training set; however, the headers in the fasta file
984 include the supraspecies taxon (AB) as an extra level between the genus (G) and species
985 taxonomic levels (A, B or AB), as illustrated here.

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

994 **Figure 4. Trimming the training set to the specific sequenced region further**
995 **reduces the error rate.** (**a**) The percentage of eHOMD-derived simulated reads classified
996 at species level using the ***FL_Compilation_TS*** (orange) training set compared to

997 subsequent trimmed versions ***V1V3_Raw_TS*** (green) and ***V1V3_Curated_TS*** (red). (**b**)
998 The percentage of classified reads that were misclassified with each of these three
999 training sets. The naïve Bayesian RDP Classifier was used with bootstrap values ranging
1000 from 50 to 100. (**c**) This graph, which is specific to the eHOMD training set construction
1001 (**V1V3_eHOMDSim_250N100** dataset), indicates how researchers can determine the
1002 bootstrap value to use with the naïve Bayesian RDP Classifier by deciding an acceptable
1003 level of the % of reads misclassified (blue line; e.g., 0.5%) and/or of the % of reads that
1004 are not classified (red line).

1005 **Figure 5. Addition of a supraspecies level to the training set increases the**
1006 **percentage of classified reads.** (**a**) The percentage of eHOMD-derived simulated reads
1007 classified at species/supraspecies level using the ***V1V3_Curated_TS*** training set (red)
1008 versus the ***FL_Supraspecies_TS*** training set (blue). (**b**) Percentage of classified reads
1009 that were misclassified with each of these TS. The naïve Bayesian RDP Classifier was
1010 used with bootstrap values ranging from 50 to 100.

1011

1012 **ADDITIONAL FILES.**

1013

1014 **Additional File 1**: IlluminaQC.pdf. Quality Phred score plots of representative 16S rRNA
1015 gene V1-V3 region Illumina runs.

1016 **Additional file 2**: V1V3_eHOMDSim_250N100.fa. The simulated eHOMD-derived
1017 **V1V3_eHOMDSim_250N100** dataset.

1018 **Additional file 3**: eHOMDv15.1_FL_Compilation_TS.fa. The 16S rRNA gene full-length
1019 eHOMD training set (***FL_Compilation_TS***).

1020 **Additional file 4**: eHOMDv15.1_V1V3_Supraspecies_TS.fa. The 16S rRNA gene V1-V3
1021 eHOMD training set (***V1V3_Supraspecies_TS***).

1022 **Additional file 5**: FL_sinonasal_SMRT_ASV.fa. Sequences of the resulting 204 ASVs in
1023 the Earl-Mell sinonasal 16S rRNA gene dataset (**FL_sinonasal_SMRT_ASV** dataset).

1024 **Additional file 6**: Earl-Mell_sinonasal_analysis.xlsx with legend and tabs A-D. Species-
1025 level taxonomic assignment of the Earl-Mell sinonasal 16S rRNA gene dataset.

1026 **Additional file 7**: V1V3_sinonasal_SMRT_ASV.fa. Sequences of the resulting 204 ASVs
1027 in the **FL_sinonasal_SMRT_ASV** dataset trimmed to the V1-V3 region
1028 (**V1V3_sinonasal_SMRT_ASV** dataset).

1029 **Additional file 8**: V1V3_hADT_CL.fa. V1-V3 trimmed 16S rRNA gene human
1030 aerodigestive tract clone library dataset (**V1V3_hADT_CL** dataset).

1031 **Additional file 9**: V1V3_HMPnares_ASV.fa. ASVs derived from the HMP nares 16S
1032 rRNA gene V1-V3 dataset (**V1V3_HMPnares_ASV** dataset).

1033 **Additional file 10**: V1V3_Primers.xlsx. Sequencing primers used to amplify the V1-V3
1034 region of the 16S rRNA gene.

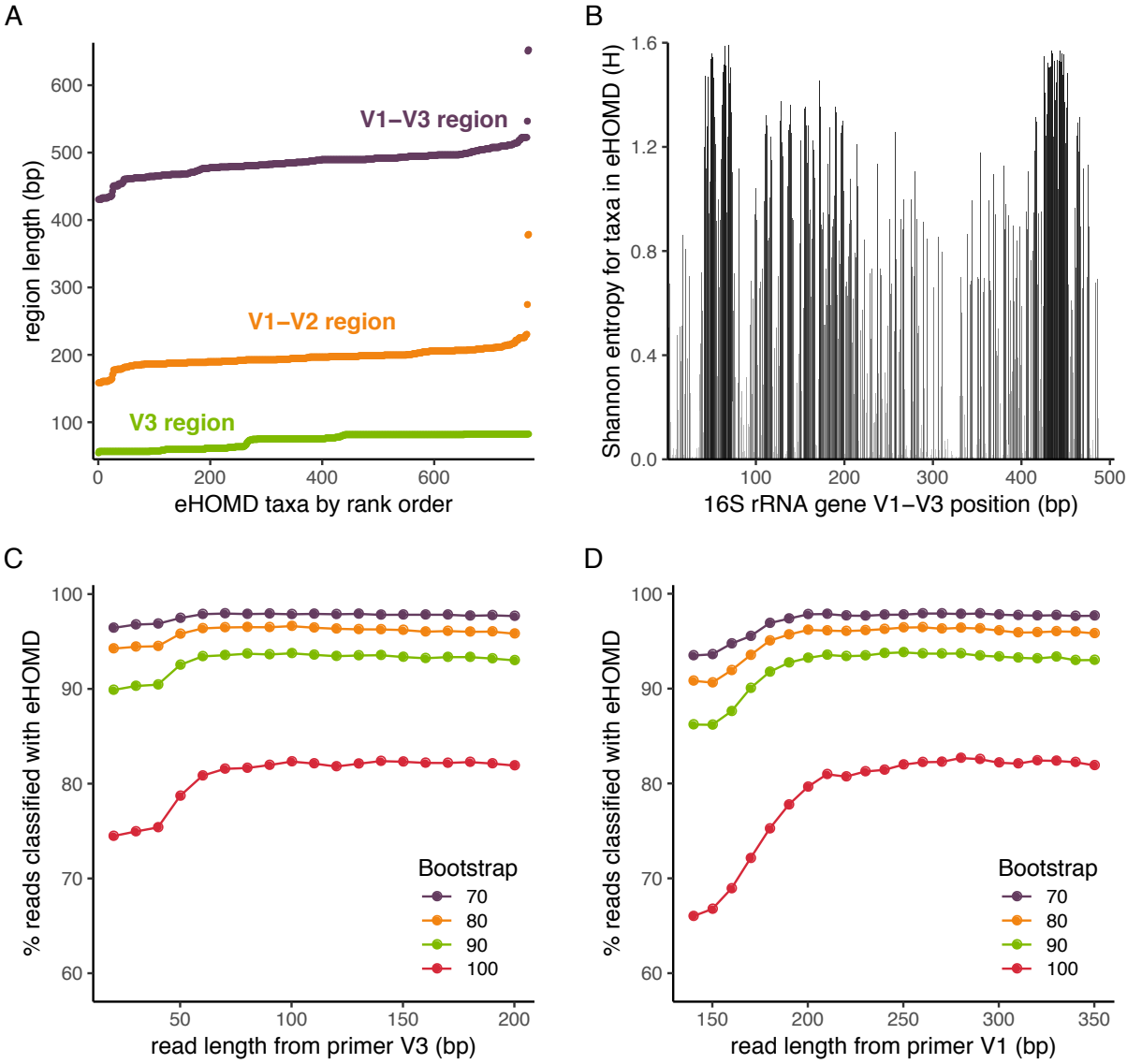
1035 **Additional file 11**: retrieveSegment.py. Custom script developed to generate a trimmed
1036 version of a training set (in our example to positions 40-880 in the gapped eHOMDrefs
1037 alignment) from a compilation of 16S rRNA gene full-length sequences.

1038 **Additional file 12**: Simulated_R2_350_R1_20-200.7z. Simulated eHOMD-derived
1039 dataset versions with R2 fixed at 350 bp and R1 fragments from the V3 primer ranging
1040 from 20 to 200 bp.

1041 **Additional file 13**: Simulated_R2_140-350_R1_200.7z. Simulated eHOMD-derived
1042 dataset versions with R1 fixed at 200 bp and R2 fragments from the V1 primer ranging
1043 from 140 to 350 bp.

1044 **Additional file 14**: FL_hADT_CL.fa. Full-length 16S rRNA gene human aerodigestive
1045 tract clone library dataset (**FL_hADT_CL** dataset).

1046 **Additional file 15**: eight2seven.R. Custom-written R function that converts the
1047 dada2::assignTaxonomy() output from eight to seven taxonomic levels for compatibility
1048 with downstream applications.


A
region length (bp)
600
500
400
300
200
100
V1–V3 region
V1–V2 region
V3 region
0
200
400
600
eHOMD taxa by rank order
B
Shannon entropy for taxa in eHOMD (H)
1.6
1.2
0.8
0.4
0.0
100
200
300
400
500
16S rRNA gene V1–V3 position (bp)
C
% reads classified with eHOMD
100
90
80
70
60
50
100
150
200
read length from primer V3 (bp)
Bootstrap
70
80
90
100
D
% reads classified with eHOMD
100
90
80
70
60
150
200
250
300
350
read length from primer V1 (bp)
Bootstrap
70
80
90
100

A

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

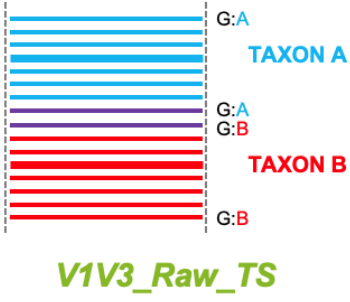
$$P(w_i|T_B) = \frac{0 + P_i}{M + 1}$$

B

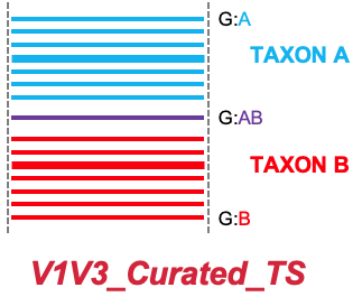
$$P(w_i|T_A) = \frac{m(w_i + P_i)}{M + 1}$$

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

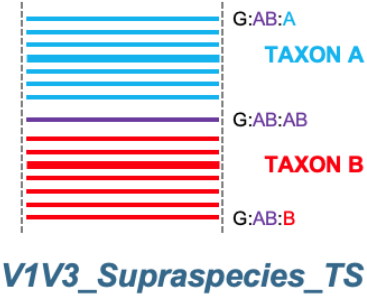
C

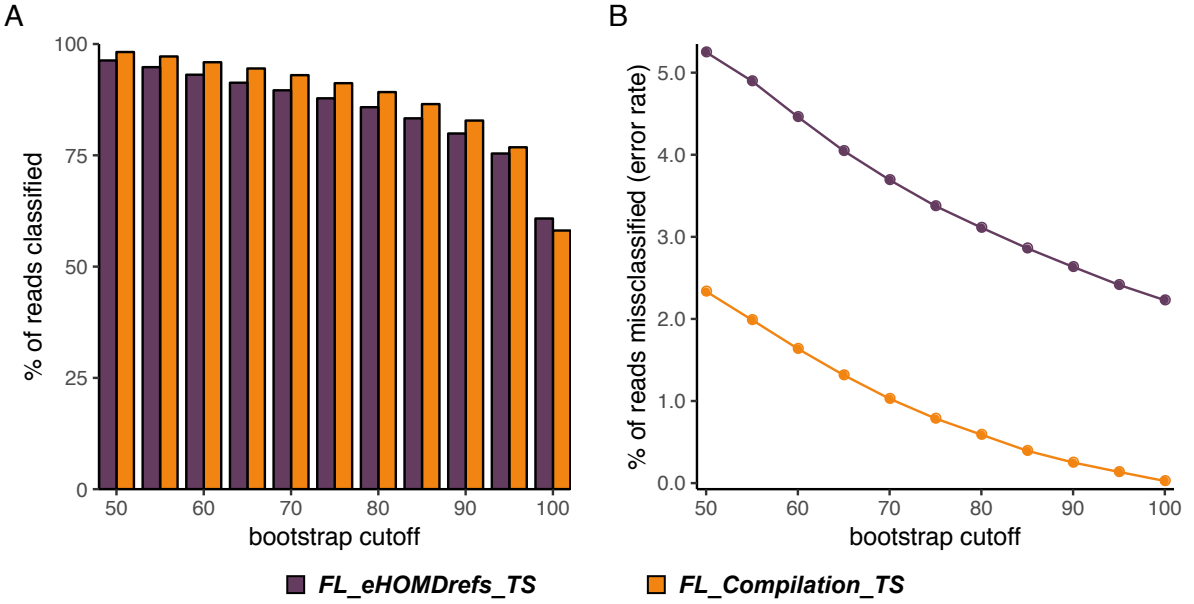


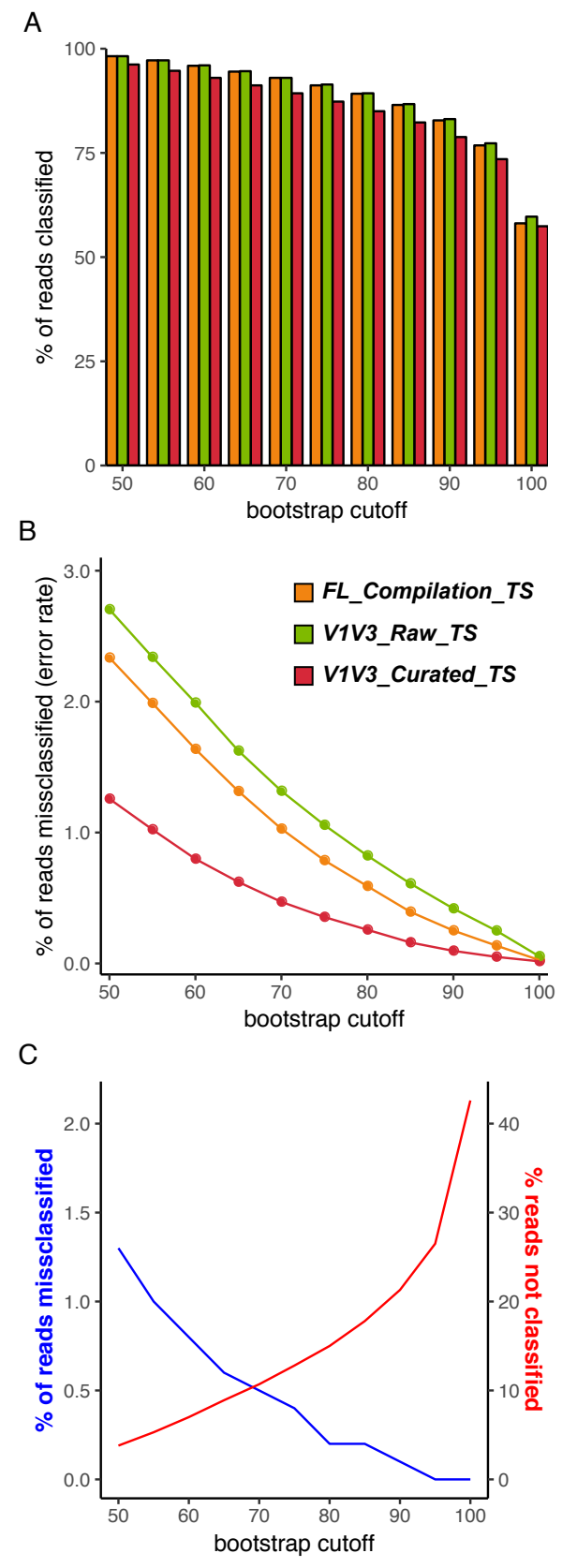
D

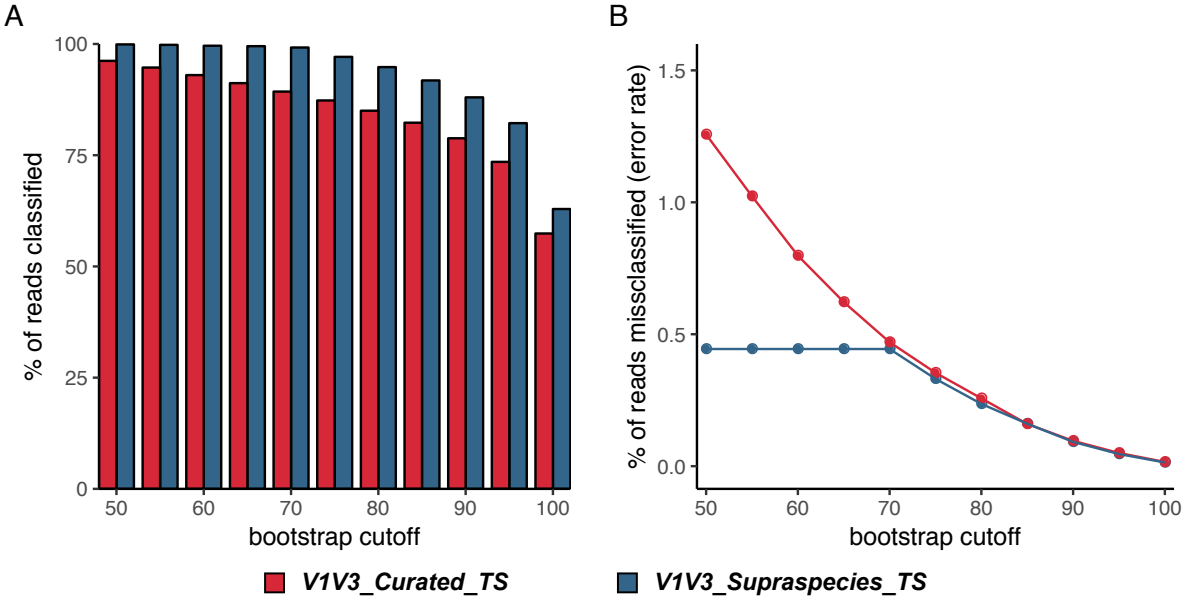


E




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


A
% of reads classified
100
75
50
25
0
50
60
70
80
90
100
bootstrap cutoff
B
% of reads missclassified (error rate)
3.0
2.0
1.0
0.0
FL_Compilation_TS
V1V3_Raw_TS
V1V3_Curated_TS
50
60
70
80
90
100
bootstrap cutoff
C
% of reads missclassified
2.0
1.5
1.0
0.5
0.0
% reads not classified
40
30
20
10
0
50
60
70
80
90
100
bootstrap cutoff


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