New insights into human nostril microbiome from the *expanded* Human Oral Microbiome Database (*e*HOMD): a resource for species-level identification of microbiome data from the aerodigestive tract

Isabel F. Escapa[a,b], Tsute Chen[a,b*], Yanmei Huang[a,b*], Prasad Gajare[a], Floyd E. Dewhirst[a,b], Katherine P. Lemon[a,c#]

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

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

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

Running Title: *e*HOMD: a respiratory tract and oral microbial database

* T.C. and Y.H. contributed equally to this work.

Address correspondence to Katherine P. Lemon, klemon@forsyth.org

key words: Microbiota; Microbiome; Metagenomics; Nasal; Nostril; Nares; Pharynx; Mouth; Esophagus; Sinus; Ribosomal, 16S; Nucleic Acid; Databases; Algorithms; Sequence Analysis; *Staphylococcus*; *Corynebacterium*; *Dolosigranulum*; *Lawsonella*

abstract word count: 250

text word count: 5534

ABSTRACT

A comprehensive microbiome database for sites along the human aerodigestive tract revealed new insights into the nostril microbiome. The *expanded* Human Oral Microbiome Database (*e*HOMD) provides well-curated 16S rRNA gene reference sequences linked to available genomes. The *e*HOMD enables assignment of species-level taxonomy to NextGeneration sequences derived from diverse aerodigestive tract sites, including the nasal passages, sinuses, throat, esophagus and mouth. Using Minimum Entropy Decomposition coupled with the RDP Classifier and our *e*HOMD V1-V3 training set, we reanalyzed 16S rRNA V1-V3 sequences from the nostrils of 210 Human Microbiome Project participants at the species level revealing four key insights. First, we discovered that *Lawsonella clevelandensis*, a recently named bacterium, and *Neisseriaceae* [G-1] HMT-174, a previously unrecognized bacterium, are common in the adult nostrils. Second, just 19 species accounted for 90% of the total sequences from all participants. Third, one of these 19 belonged to a currently uncultivated genus. Fourth, for 94% of the participants, 2 to 10 species constituted 90% of their sequences, indicating that nostril microbiome may be represented by limited consortia. These insights highlight the strengths of the nostril microbiome as a model system for studying interspecies interactions and microbiome function. Also, in this cohort, four common nasal species (*Dolosigranulum pigrum* and three *Corynebacterium* species) showed positive differential abundance when the pathobiont *Staphylococcus aureus* was absent, generating hypotheses regarding colonization resistance. By facilitating species-level taxonomic assignment to microbes from the human aerodigestive tract,

the *e*HOMD is a vital resource enhancing clinical relevance of 16S rRNA gene-based microbiome studies.

IMPORTANCE

The *e*HOMD (ehomd.org) is a valuable resource for researchers, from basic to clinical, who study the microbiomes in health and disease of body sites in the human aerodigestive tract, including the nasal passages, sinuses, throat, esophagus and mouth, and the lower respiratory tract. The *e*HOMD is an actively curated, web-based, open-access resource. *e*HOMD provides the following: (1) species-level taxonomy based on grouping 16S rRNA gene sequences at 98.5% identity, (2) a systematic naming scheme for unnamed and/or uncultivated microbial taxa, (3) reference genomes to facilitate metagenomic and metatranscriptomic studies and (4) convenient cross-links to other databases (e.g., PubMed and Entrez). By facilitating the assignment of species names to sequences, the *e*HOMD is a vital resource for enhancing the clinical relevance of 16S rRNA gene-based microbiome studies.

INTRODUCTION

The human aerodigestive tract includes the oral cavity, pharynx, esophagus, nasal passages and sinuses. Because the aerodigestive tract microbiome commonly harbors both harmless and pathogenic bacterial species of the same genus, optimizing the clinical relevance of microbiome studies for body sites within the aerodigestive tract requires sequence identification at the species or, at least, subgenus level. Understanding the composition and function of the microbiome of the aerodigestive tract

is important for understanding human health and disease since aerodigestive tract sites are often colonized by common bacterial pathogens and are associated with prevalent diseases characterized by dysbiosis.

The reductions in the cost of NextGeneration DNA Sequencing (NGS) combined with the increasing ease of determining bacterial community composition using short NGS-generated 16S rRNA gene fragments now make this a practical approach for molecular epidemiological, clinical and translational studies (1). Optimal clinical relevance of such studies requires species-level identification (2); however, to date, 16S rRNA gene-tag studies of the human microbiome are overwhelmingly limited to genus-level resolution. For example, many studies of nasal microbiota fail to distinguish medically important pathogens, e.g., *Staphylococcus aureus*, from generally harmless members of the same genus, e.g., *Staphylococcus epidermidis*. With this expansion, the *e*HOMD becomes a comprehensive web-based resource for NGS microbiome analyses of body sites throughout the aerodigestive tract in both health and disease (ehomd.org). The *e*HOMD should also serve as an effective resource for lower respiratory tract (LRT) microbiome studies based on the breadth of taxa included, and that many LRT microbes are found in the mouth, pharynx and nasal passages (3).

For many bacterial taxa, newer computational methods, e.g., Minimum Entropy Decomposition (MED), an unsupervised form of oligotyping (4), and DADA2 (5), parse NGS-generated short 16S rRNA gene sequences to species-level, sometimes strain-level, resolution. However, to achieve species-level taxonomy assignment for the resulting oligotypes/phylotypes, these methods must be used in conjunction with a high-resolution 16S rRNA gene taxonomic database and a classifying algorithm. Similarly,

metagenomic sequencing provides species- and, often, strain-level resolution when coupled with a reference dataset that includes genomes from multiple strains for each species. For the mouth, the HOMD (6, 7) has enabled analysis/reanalysis of oral 16S rRNA gene short-fragment datasets with these new computational tools, revealing microbe-microbe and host-microbe species-level relationships (8-10), and has been a resource for easy access to genomes from which to build reference sets for metagenomic and metatranscriptomic studies. We have also considerably expanded the number of genomes linked to aerodigestive tract taxa. Thus, the *e*HOMD now enables the broad community of researchers studying the nasal passages, sinuses, throat, esophagus and mouth to leverage newer high-resolution approaches to study the microbiome of these aerodigestive tract body sites.

The *e*HOMD also facilitates rapid comparison of 16S rRNA gene sequences from studies worldwide by providing a systematic provisional naming scheme for unnamed taxa identified through sequencing (7). Each high-resolution taxon in *e*HOMD, as defined by 98.5% sequence identity across close-to-full-length 16S rRNA gene sequences, is assigned a unique human microbial taxon number, an HMT, which can be used to search and retrieve that sequence-based taxon from any sequence dataset or database.

Here, in section I, we describe the process of generating the *e*HOMDv15.1 (ehomd.org), its utility using both 16S rRNA gene clone library and short-read datasets and, in section II, new discoveries about the nostril microbiome based on analysis using the *e*HOMD.

RESULTS and DISCUSSION

I. The *e*HOMD is a Resource for Microbiome Research on the Human Upper Digestive and Respiratory Tracts.

As described below, the *e*HOMD is a comprehensive, actively curated, web-based resource open to the entire scientific community that classifies 16S rRNA gene sequences at a high resolution (98.5% sequence identity); provides a systematic provisional naming scheme for as-yet unnamed/uncultivated taxa and provides a resource for easily searching available genomes for included taxa. The *e*HOMD facilitates the identification of aerodigestive and lower respiratory tract bacteria and provides phylogenetic, genomic, phenotypic, clinical and bibliographic information for these microbes.

The *e*HOMD outperforms other 16S rRNA gene databases for taxonomic identification of 16S rRNA gene clones from the human nostrils. Here we describe the generation of *e*HOMDv15.1, which outperformed three other commonly used 16S rRNA gene databases in assigning species-level taxonomy to sequences in a dataset of nostril-derived 16S rRNA gene clones (Table 1). Species-level taxonomy assignment was defined as 98.5% identity with 98% coverage via blastn. An initial analysis showed that HOMDv14.5 enabled species-level taxonomic assignment to only 50.2% of the 44,374 16S rRNA gene clones from nostril (anterior nares) samples generated by Julie Segre, Heidi Kong and colleagues, henceforth the SKn dataset (Table 1) (11-16). To expand HOMD to be a resource for the microbiomes of the entire human aerodigestive tract, we started with nasal- and sinus-associated bacterial species. As illustrated in Figure 1, and described in detail in the methods, we compiled a list of candidate nasal and sinus species gleaned from culture-dependent studies (17-19) plus anaerobes

cultivated from cystic fibrosis sputa (20) (Table S1). To assess which of these candidate species are most likely to be common members of the nasal microbiome, we used blastn to identify those present in the SKn dataset. We then added these to a provisional expanded database (Fig. 1, A). Using blastn, we assayed how well this provisional *e*HOMDv15.01 captured clones in the SKn dataset (Table S2). Examination of sequences in the SKn dataset that were not identified resulted in further additions generating the provisional *e*HOMDv15.02 (Fig. 1, B and C). Next, we evaluated how well *e*HOMDv15.02 served to identify sequences in the SKn clone dataset using blastn (Fig. 1, D). We then took an iterative approach using blastn to evaluate the performance of *e*HOMDv15.02 against a set of three V1-V2 or V1-V3 16S rRNA gene short-read datasets (21-24) and two close-to-full-length 16S rRNA gene clone datasets from the aerodigestive tract in children and adults in health and disease (25, 26) in comparison to three commonly used 16S rRNA gene databases: NCBI 16S Microbial (NCBI 16S) (27), RDP16 (28) and SILVA128 (29, 30) (Fig. 1, E and Table S3). These steps resulted in the generation of the provisional *e*HOMDv15.03. Further additions to include taxa that can be present on the skin of the nasal vestibule (nostril or nares samples) but which are more common at other skin sites resulted from using blastn to analyze the full Segre-Kong skin 16S rRNA gene clone dataset, excluding nostrils, (the SKs dataset) (11-16) against both *e*HOMDv15.03 and SILVA128 (Fig. 1, F and G). Based on these results, we generated the *e*HOMDv15.1, which identified 95.1% of the 16S rRNA gene reads in the SKn dataset outperforming the three other commonly used 16S rRNA gene databases (Table 1). Each step in this process improved *e*HOMD with respect to

158 identification of clones from the SKn dataset, establishing *e*HOMD as a resource for the
159 human nasal microbiome (Fig. 1 and Table S2).
160 **The *e*HOMD is a resource for taxonomic assignment of 16S rRNA gene**
161 **sequences from the entire human aerodigestive tract, as well as the lower**
162 **respiratory tract.** To assess the value of *e*HOMDv15.1 for analysis of datasets from
163 sites throughout the human aerodigestive tract, it was compared with three commonly
164 used 16S rRNA gene databases and consistently performed better than or comparable
165 to these databases (Table 2). For these comparisons, we used blastn to assign
166 taxonomy to three short-read (V1-V2 and V1-V3) and five approximately full-length-
167 clone-library 16S rRNA gene datasets that are publicly available (21-23, 25, 26, 31-33).
168 For short-read datasets, we focused on those covering all or part of the V1-V3 region of
169 the 16S rRNA gene for three reasons. First, the V1-V3 region is preferentially used for
170 the mouth. Second, for skin, V1-V3 sequencing results show high concordance with
171 those from metagenomic sequencing (34). Third, to enable species-level distinctions
172 within respiratory tract genera that include both common commensals and pathogens,
173 V1-V3 is preferable for the nasal passages, sinuses and nasopharynx (2, 35-37). The
174 chosen datasets include samples from children or adults in health and/or disease. The
175 samples in these datasets are from human nostril swabs (21, 23), nasal lavage fluid
176 (22), esophageal biopsies (25), extubated endotracheal tubes (32), endotracheal tube
177 aspirates (31), sputa (33) and bronchoalveolar lavage (BAL) fluid (26). Endotracheal
178 tube sampling may represent both upper and lower respiratory tract microbes and
179 sputum may be contaminated by oral microbes, whereas BAL fluid represents microbes
180 present in the lower respiratory tract. Therefore, these provide broad representation for

bacterial microbiota of the human aerodigestive tract, as well as the human lower respiratory tract (Table 2). The composition of the bacterial microbiota from the nasal passages varies across the span of human life (1) and *e*HOMD captures this variability. Similar to *e*HOMD, SILVA128 is also actively curated (29, 30). However, SILVA128 includes bacteria from a much larger range of habitats, whereas, *e*HOMD is focused on human-associated bacteria from the respiratory and upper digestive tracts. The performance of *e*HOMDv15.1 establishes it as a resource for microbiome studies of all body sites within the human respiratory and upper digestive tracts.

The *e*HOMDv15.1 performed very well for nostril samples (Tables 1 and 2), which are a type of skin microbiome sample since the nostrils open onto the skin-covered surface of the nasal vestibules. Based on this, we hypothesized that *e*HOMD might also perform well for other skin sites. To test this, we used *e*HOMDv15.04 to perform blastn for taxonomic assignment of 16S rRNA gene reads from the complete set of clones from multiple nonnasal skin sites generated by Segre, Kong and colleagues (SKs dataset) (11-16). As shown in Table 3, *e*HOMDv15.04 performed very well for oily skin sites (alar crease, external auditory canal, back, glabella, manubrium, retroauricular crease and occiput) and the nostrils (nares), identifying >88% of the clones, which was more than the other databases for six of these eight sites. Either SILVA128 or *e*HOMDv15.04 consistently identified the most clones for each skin site to species level (98.5% identity and 98% coverage); *e*HOMDv15.04 is almost identical to the released *e*HOMDv15.1. In contrast, *e*HOMDv15.04 performed less well than SILVA128 for the majority of the moist skin sites (Table 3), e.g., the axillary vault (arm pit). A review of the details of these

203 results revealed that a further expansion comparable to what we did here is necessary
204 for *e*HOMD to include the full diversity of all skin sites.
205 **The *e*HOMD is a resource for annotated genomes matched to HMTs for use in**
206 **metagenomic and metatranscriptomic studies.** Well-curated and annotated
207 reference genomes correctly named at the species level are a critical resource for
208 mapping metagenomic and metatranscriptomic data to gene and functional information,
209 and for identifying species-level activity within the microbiome. There are currently
210 >160,000 microbial genomic sequences deposited to GenBank; however, many of these
211 genomes remain poorly or not-yet annotated or lack species-level taxonomy
212 assignment, thus limiting the functional interpretation of
213 metagenomic/metatranscriptomic studies to the genus level. Therefore, one goal of the
214 *e*HOMD is to provide correctly named, curated and annotated genomes for all HMTs;
215 this is an ongoing process. In generating *e*HOMDv15.1, we determined the species-
216 level assignment for 112 genomes in GenBank that were previously identified only to
217 the genus level and which matched to 27 *e*HOMD taxa. For each of these genomes, the
218 phylogenetic relationship to the assigned HMT was verified by both phylogenetic
219 analysis using 16S rRNA gene sequences and by phylogenomic analysis using a set of
220 core proteins and PhyloPhlAn (38). To date, 85% (475) of the cultivated taxa (and 62%
221 of all taxa), included in *e*HOMD have at least one sequenced genome.
222 **The *e*HOMD is a resource for species-level assignment to the outputs of high-**
223 **resolution 16S rRNA gene analysis algorithms.** Newer algorithms, such as DADA2
224 and MED, permit high-resolution binning of 16S rRNA gene short-read sequences (4,
225 5). Moreover, the RDP naïve Bayesian Classifier is an effective tool for assigning

taxonomy to 16S rRNA gene sequences, both full length and short read, when coupled with a robust, well-curated training set (39, 40). Together these tools permit species-level analysis of short-read 16S rRNA gene datasets. Because the V1-V3 region is the most informative short-read fragment for most of the common bacteria of the aerodigestive tract, we generated a training set for the V1-V3 region of the 16S rRNA gene that includes all taxa represented in the *e*HOMD, which is described elsewhere. In our training set, we grouped taxa that were indistinguishable based on the sequence of their V1-V3 region together as supraspecies to preserve subgenus-level resolution, e.g., *Staphylococcus capitis_caprae.*

Advantages and limitations of the *e*HOMD. The *e*HOMD has advantages and limitations when compared to other 16S rRNA gene databases, such as RDP, NCBI, SILVA and Greengenes (27-30, 41). Its primary distinction is that *e*HOMD is dedicated to providing taxonomic, genomic, bibliographic and other information specifically for the approximately 800 microbial taxa found in the human aerodigestive tract. Here, we highlight four advantages of *e*HOMD. First, the *e*HOMD is based on extensively curated 16S rRNA reference sets (*e*HOMDrefs) and a taxonomy that uses phylogenetic position in 16S rRNA-based trees rather than a taxon's currently assigned, or misassigned, taxonomic name (7). For example, the genus "*Eubacteria*" in the phylum Firmicutes includes members that should be divided into multiple genera in 7 different families (42). In *e*HOMD, members of the "*Eubacteria"* are placed in their phylogenetically appropriate family, e.g., *Peptostreptococcaceae*, rather than incorrectly into the family *Eubacteriaceae.* Appropriate taxonomy files are readily available from *e*HOMD for mothur (43) and other programs. Second, because *e*HOMD includes a provisional

249 species-level naming scheme, sequences that can only be assigned genus-level
250 taxonomy in other databases are resolved to species level via an HMT number. This
251 enhances the ability to identify and learn about taxa that currently lack full identification
252 and naming. Third, in *e*HOMD, for the 475 taxa with at least one sequenced genome,
253 genomes can be viewed graphically or searched using blastn, blastp or tblastn. For taxa
254 lacking available genomic sequences the available 16S rRNA sequences are included.
255 Many genomes of aerodigestive tract organisms are in the whole-genome shotgun
256 contigs (wgs) section of NCBI and are visible by blast search only through wgs provided
257 that you know the genome and can provide the BioProjectID or WGS Project ID. At
258 *e*HOMD, one can readily compare dozens to over a hundred genomes for some taxa to
259 begin to understand the pangenome of aerodigestive tract microbes. Fourth, we have
260 also complied proteome sequence sets for genome-sequenced taxa enabling
261 proteomics and mass spectra searches on a dataset limited to proteins from ~2,000
262 relevant genomes.
263 In terms of limitations, the taxa included in the *e*HOMD, the 16S rRNA reference
264 sequences and genomes are not appropriate for samples from 1) human body sites
265 outside of the aerodigestive and respiratory tracts, 2) nonhuman hosts or 3) the
266 environment. In contrast, RDP (28), SILVA (29, 30) and Greengenes (41) are curated
267 16S rRNA databases inclusive of all sources and environments. Whereas, the NCBI
268 16S database is a curated set of sequences for bacterial and archaeal named species
269 only (a.k.a RefSeqs) that is more frequently updated (27). Finally, the NCBI nucleotide
270 database (nr/nt) includes the largest set of 16S rRNA sequences available; however,
271 the vast majority have no taxonomic attribution and are listed as simply “uncultured

bacterium clone." Thus, RDP, SILVA, NCBI, Greengenes and other similar general databases have advantages for research on microbial communities outside the human respiratory and upper digestive tracts, whereas *e*HOMD is preferred for the microbiomes of the human upper digestive and respiratory tracts.

II. The *e*HOMD revealed previously unknown properties of the human nasal microbiome.

To date the human nasal microbiome has mostly been characterized at the genus level. For example, the Human Microbiome Project (HMP) characterized the bacterial community in the adult nostrils (nares) to the genus level using 16S rRNA sequences (23, 24). However, the human nasal passages can host a number of genera that include both common commensals and important bacterial pathogens, e.g., *Staphylococcus, Streptococcus, Haemophilus, Moraxella* and *Neisseria* (review in (1)). Thus, species-level nasal microbiome studies are needed from both a clinical and ecological perspective. Therefore, to further our understanding of the adult nostril microbiome, we used MED, the RDP classifier and our *e*HOMD V1-V3 training set to reanalyze a subset of the HMP nares V1-V3 16S rRNA dataset consisting of one sample each from 210 adults. Henceforth, we refer to this subset as the HMP nares V1-V3 dataset. This resulted in species/supraspecies-level taxonomic assignment for 95% of the sequences and revealed new insights into the adult nostril microbiome, which are described below.

A small number of cultivated species account for the majority of the adult nostril microbiome. Genus-level information from the HMP corroborates data from smaller cohorts showing the nostril microbiome has a very uneven distribution both overall and per person, reviewed in (44). In our reanalysis, 10 genera accounted for 95% of the total

295 reads from 210 adults, with the remaining genera each present at very low relative
296 abundance and prevalence (Fig. 2A and Table S4). Moreover, for the majority of
297 participants, 5 or fewer genera constituted 90% of the sequences in their sample (Fig.
298 2B). This uneven distribution characterized by the numeric dominance of a small
299 number of taxa was even more striking at the species level (45). We found that the 6
300 most relatively abundant species made up 72% of the total sequences, and the top 5
301 each had a prevalence of ≥81% (Fig. 2C and Table S5). Moreover, between 2 and 10
302 species accounted for 90% of the sequences in 94% of the participants (Fig. 2D). Also,
303 just 19 species/supraspecies-level taxa constituted 90% of the total 16S rRNA gene
304 sequences from all 210 participants together (Table S5), and only one of these
305 belonged to an as-yet-uncultivated genus, as described below. The implication of these
306 findings is that *in vitro* consortia consisting of small numbers of species can effectively
307 represent the natural nasal community, facilitating functional studies of the nostril
308 microbiome.

309 **Identification of two previously unrecognized common nasal bacterial taxa**.
310 Reanalysis of both the HMP nares V1-V3 dataset and the SKn 16S rRNA gene clone
311 dataset revealed two previously unrecognized taxa are common in the nostril
312 microbiome: *Lawsonella clevelandensis* and an unnamed *Neisseriaceae* [G-1]
313 bacterium, to which we assigned the provisional name *Neisseriaceae* [G-1] bacterium
314 HMT-174. These are discussed in further detail below.

315 **The human nasal passages are the primary habitat for a subset of bacterial**
316 **species.** The topologically external surfaces of the human body are the primary habitat
317 for a number of bacterial taxa, which are often present at both high relative abundance

318 and high prevalence in the human microbiome. The previous HOMDv14.5 featured
319 putative body site habitat assignment for the majority of taxa included. In generating the
320 *e*HOMDv15.1, we hypothesized that comparing the relative abundance of sequences
321 identified to species or supraspecies level in the SKn clones and the SKs clones
322 (nonnasal skin sites) would permit putative identification of the primary body-site habitat
323 for a subset of nostril-associated bacteria. Based on criteria described in the methods,
324 we identified 13 species as having the nostrils as their primary habitat (Table S6).
325 Online at http://ehomd.org/index.php?name=HOMD the primary body site for each
326 taxon is denoted as oral, nasal, skin, vaginal or unassigned. Definitive identification of
327 the primary habitat of all human-associated bacteria will require species-level
328 identification of bacteria at each distinct habitat across the surfaces of the human body
329 from a cohort of individuals. This would enable a more complete version of the type of
330 comparison performed here.
331 Members of the genus *Corynebacterium* (phylum Actinobacteria) are common in human
332 nasal, skin and oral microbiomes but their species-level distribution across these body
333 sites remains less clear (23). Our analysis of the SKns clones identified three
334 *Corynebacterium* as primarily located in the nostrils compared to the other skin sites: *C.*
335 *propinquum, C. pseudodiphtheriticum* and *C. accolens* (Table S6). In the species-level
336 reanalysis of the HMP nares V1-V3 dataset, these were among the top five
337 *Corynebacterium* species/supraspecies by rank order abundance of sequences (Table
338 S5). In this reanalysis, *Corynebacterium tuberculostearicum* accounted for the fourth
339 largest number of sequences; however, in the SKns clones it was not disproportionately
340 present in the nostrils. Therefore, although common in the nostrils, we did not consider

the nostrils the primary habitat for *C. tuberculostearicum*, in contrast to *C. propinquum, C. pseudodiphtheriticum* and *C. accolens.*

The human skin and nostrils are primary habitats for *Lawsonella clevelandensis*.

In 2016, *Lawsonella clevelandensis* was described as a novel genus and species within the suborder *Corynebacterineae* (phylum *Actinobacteria*) (46); genomes for two isolates are available (47). It was initially isolated from several human abscesses, mostly from immunocompromised hosts, but its natural habitat was unknown. This led to speculation *L. clevelandensis* might either be a member of the human microbiome or an environmental microbe with the capacity for opportunistic infection [Harrington, 2013 #160;Bell, 2016 #1601]. Our results indicate that *L. clevelandensis* is a common member of the bacterial microbiome of some oily skin sites and the nostrils of humans (Table S6). Indeed, in the SKn clones, we detected *L. clevelandensis* as the 11th most abundant taxon. Validating the SKn data in our reanalysis of the HMP nares V1-V3 dataset from 210 participants, we found that *L. clevelandensis* was the 5th most abundant species overall with a prevalence of 86% (Table S5). In the nostrils of individual HMP participants, *L. clevelandensis* had an average relative abundance of 5.7% and a median relative abundance of 2.6% (range 0 to 42.9%). *L. clevelandensis* is recently reported to be present on skin (48). Our reanalysis of the SKns clones indicated that of these body sites the primary habitat for *L. clevelandensis* is oily skin sites, in particular the alar crease, glabella and occiput where it accounts for higher relative abundance than in the nostrils (Table S6). Virtually nothing is known about the role of *L. clevelandensis* in the human microbiome. By report, it grows best under anaerobic conditions (<1% O_2) and cells are a mixture of pleomorphic cocci and bacilli

that stain gram-variable to gram-positive and partially acid fast (46, 47). Based on its 16S rRNA gene sequence, *L. clevelandensis* is most closely related to the genus *Dietzia*, which includes mostly environmental species. Within its suborder *Corynebacterineae* are other human associated genera, including *Corynebacterium*, which is commonly found on oral, nasal and skin surfaces, and *Mycobacterium.* Our analyses demonstrate *L. clevelandensis* is a common member of the human skin and nasal microbiomes, opening up opportunities for future research on its ecology and its functions with respect to humans.

The majority of the bacteria detected from the human nasal passages are cultivated. Using blastn to compare the 16S rRNA gene SKn clones with *e*HOMDv15.1, we found that 93.1% of these sequences from adult nostrils can be assigned to cultivated named species, 2.1% to cultivated unnamed taxa, and 4.7% to uncultivated unnamed taxa. In terms of the total number of species-level taxa represented by the SKn clones, rather than the total number of sequences, 70.1% matched to cultivated named taxa, 14.4% to cultivated unnamed taxa, and 15.5% uncultivated unnamed taxa. Similarly, in the HMP nares V1-V3 dataset from 210 participants (see below), 91.1% of sequences represented cultivated, named bacterial species. Thus, the bacterial microbiota of the nasal passages is numerically dominated by cultivated bacteria. In contrast, approximately 30% of the oral microbiota (ehomd.org) and a larger, but not precisely defined, fraction of the intestinal microbiota is currently uncultivated (49). The ability to cultivate the majority of species detected in the nasal microbiota is an advantage when studying the functions of members of the nasal microbiome.

One common nasal taxon remains to be cultivated. In exploring the SKn dataset to generate *e*HOMD, we realized that the 12th most abundant clone in the SKn dataset lacked genus-level assignment. It represents an unnamed and apparently uncultivated *Neisseriaceae* bacterial taxon to which we have assigned the provisional name *Neisseriaceae* [G-1] bacterium HMT-174. Its provisional naming facilitates recognition of this bacterium in other datasets and its future study. In our reanalysis of the HMP nares V1-V3 dataset, *Neisseriaceae* [G-1] bacterium HMT-174 was the 10th most abundant species overall with a prevalence of 35%. In individual participants, it had an average relative abundance of 1.3% and a median relative abundance of 0 (range 0 to 38.4%). Blastn analysis of our reference sequence for *Neisseriaceae* [G-1] bacterium HMT-174 against the 16S ribosomal RNA sequences database at NCBI gave matches of 90% to 92% similarity to members of the family *Neisseriaceae* and matches to the neighboring family *Chromobacteriaceae* at 88% to 89%. A phylogenetic tree of taxon HMT-174 with members of these two families was more instructive since it clearly placed taxon HMT-174 as a deeply branching, but monophyletic, member of the *Neisseriaceae* family with the closest named taxa being *Snodgrassella alvi* (NR_118404) at 92% similarity and *Vitreoscilla stercoraria* (NR_0258994) at 91% similarity, and the main cluster of *Neisseriaceae* at or below 92% similar (Fig. S1). The main cluster of genera in a tree of the family *Neisseriaceae* includes *Neisseria, Alysiella, Bergeriella, Conchiformibius, Eikenella, Kingella* and other mammalian host-associated taxa. There is a separate clade of the insect associated genera *Snodgrassella* and *Stenoxybacter*, whereas *Vitreoscilla* is from cow dung and forms its own clade. Recognition of the as-yet-uncultivated *Neisseriaceae* [G-1] bacterium HMT-174 as a common member of the

409 adult nostril microbiome supports future research to cultivate and characterize this
410 bacterium. *Neisseriaceae* [G-1] bacterium HMT-327 is another uncultivated nasal taxon,
411 likely from the same unnamed genus, and the 20th (HMP) and 46th (SKn) most common
412 nasal organism in the two datasets we reanalyzed. There are several additional
413 uncultured nasal bacteria in *e*HOMD, highlighting the need for sophisticated cultivation
414 studies even in the era of NGS studies. Having 16S rRNA reference sequences tied to
415 the provisional taxonomic scheme in *e*HOMD allows targeted efforts to culture the
416 previously uncultivated based on precise 16S rRNA identification methods.

417 **Species that are differentially abundant with respect to *Neisseriaceae* [G-1]**
418 **bacterium HMT-174 or *L. clevelandensis***. There is a lack of knowledge about
419 potential relationships between the two newly recognized members of the nostril
420 microbiome, *L. clevelandensis* and *Neisseriaceae* [G-1] bacterium HMT-174, and other
421 known members of the nostril microbiome. Therefore, we performed LEfSe analysis on
422 samples grouped based on the presence or absence of sequences of each of these two
423 taxa of interest in search of species displaying differential relative abundance based on
424 either of these two (50). For *Neisseriaceae* [G-1] bacterium HMT-174, this was targeted
425 at identifying potential growth partners for this as-yet-uncultivated bacterium. In the
426 presence of *Neisseriaceae* [G-1] HMT-174, *Lachnospiraceae* [G-9] bacterium HMT-924
427 showed positive differential abundance (Fig. 3A). It had a low prevalence of 1.4% and
428 low total relative abundance of 0.02% (Table S5). However, it also appears to be
429 currently uncultivated. In contrast, a common nasal bacterium, *D. pigrum* showed
430 positive differential abundance in the absence of *L. clevelandensis*, whereas no other
431 species showed differential abundance in the presence of *L. clevelandensis* (Fig. 3B

dark gray bars). Of note, *D. pigrum* is positively correlated with the genus *Corynebacterium* in a number of pediatric nasal microbiome studies, reviewed in (1). Both *Corynebacterium* and *L. clevelandensis* are members of the order *Corynebacteriales*. However, their differential relationship with respect to *D. pigrum* hints at potentially interesting differences in the biology of these two closely related genera that are both common constituents of the adult nostril microbiome.

Several common species of nasal bacteria are more abundant when *S. aureus* is absent. Finally, as proof of principle that *e*HOMD enhances the clinical relevance of 16S rRNA gene-based microbiome studies, we turned our attention to *S. aureus. S. aureus* is both a common member of the nasal microbiome and an important human pathogen, with >10,000 attributable deaths/year in the U.S. (51, 52). The genus *Staphylococcus* includes many human commensals hence the clinical importance of distinguishing *aureus* from non-*aureus* species. In our reanalysis of the HMP nares V1-V3 dataset, *S. aureus* sequences accounted for 3.9% of the total with a prevalence of 34%, consistent with it being common in the nasal microbiome (2, 53). *S. aureus* nostril colonization is a risk factor for invasive infection at distant body sites (51, 54). Therefore, in the absence of an effective vaccine (55, 56), there is increasing interest in identifying members of the nostril and skin microbiome that might play a role in colonization resistance to *S. aureus*, e.g., (57-60). Although differential abundance does not indicate causation, identifying such relationships at the species level in a cohort the size of the HMP can arbitrate variations among findings in smaller cohorts and generate new hypotheses for future testing. Therefore, we used LEfSe (50) to identify taxa displaying differential relative abundance in HMP nostril samples in which 16S rRNA

455 gene sequences corresponding to *S. aureus* were absent or present. Two of the five
456 most abundant genera in the adult nostrils, *Corynebacterium* and *Dolosigranulum* (Fig.
457 2A and Table S4), displayed the greatest positive differential abundance in the absence
458 of *S. aureus* (Fig. 3C dark gray bars). As previously reviewed (44), there is variability
459 between studies with smaller cohorts with respect to the reported correlations between
460 *S. aureus* and specific *Corynebacterium* species in the nostril microbiome; this
461 variability might relate to strain-level differences and/or to the small cohort sizes. In this
462 HMP cohort of 210 adults, 3 *Corynebacterium* species/supraspecies—*accolens*,
463 *accolens_macginleyi_tuberculostearicum* and *propinquum*—showed positive differential
464 abundance in the absence of *S. aureus* nostril colonization (Fig. 3C dark gray bars).
465 These 3 were among the 9 most abundant species in the cohort overall (Fig. 2C and
466 Table S5). Similarly, *D. pigrum* (61) showed a positive differential abundance in the
467 absence of *S. aureus* (Fig. 3C dark gray bars). This is consistent with observations from
468 Liu, Andersen and colleagues that high-levels of *D. pigrum* are the strongest predictor of
469 absence of *S. aureus* nostril colonization in 90 older adult Danish twin pairs (62). In our
470 reanalysis of the HMP nares V1-V3 dataset, *D. pigrum* was the 6th most abundant
471 species overall with a prevalence of 41% (Fig. 2C and Table S5). In contrast, the
472 species with positive differential abundance when *S. aureus* was present (Fig. 3C light
473 gray bars) all had a cumulative relative abundance of ≤0.01% and a prevalence of
474 ≤3.33%, suggesting these are unlikely to be biologically relevant in most people
475 colonized by *S. aureus* (Table S5).

476 **Summary.** As demonstrated here, the *e*HOMD (ehomd.org) is a comprehensive well-
477 curated online database for the bacterial microbiome of the entire aerodigestive tract

478 enabling species/supraspecies-level taxonomic assignment to full-length and V1-V3
479 16S rRNA gene sequences and including correctly assigned, annotated available
480 genomes. In generating the *e*HOMD, we identified two previously unrecognized
481 common members of the adult human nostril microbiome, opening up new avenues for
482 future research. As illustrated using the adult nostril microbiome, *e*HOMD can be
483 leveraged for species-level analyses of the relationship between members of the
484 aerodigestive tract microbiome, enhancing the clinical relevance of studies and
485 generating new hypotheses about interspecies interactions and the functions of
486 microbes within the human microbiome. The *e*HOMD provides a broad range of
487 microbial researchers, from basic to clinical, a resource for exploring the microbial
488 communities that inhabit the human respiratory and upper digestive tracts in health and
489 disease.
490

491 **MATERIALS AND METHODS**

492 **Generating the provisional *e*HOMDv15.01 by adding bacterial species from**
493 **culture-dependent studies.** To identify candidate Human Microbial Taxa (cHMTs), we
494 reviewed two studies that included cultivation of swabs taken from along the nasal
495 passages in both health and chronic rhinosinusitis (CRS) (18, 19) and one study of
496 mucosal swabs and nasal washes only in health (17). We also reviewed a culture-
497 dependent study of anaerobic bacteria isolated from cystic fibrosis (CF) sputa to identify
498 anaerobes that might be present in the nasal passages/sinuses in CF (20). Using this
499 approach, we identified 162 cHMTs, of which 65 were present in HOMDv14.51 and 97
500 were not (Fig. 1, A and Table S1). For each of these 97 named species, we downloaded

501 at least one 16S rRNA gene RefSeq from NCBI 16S (via a search of BioProjects 33175
502 and 33317) (27) and assembled these into a reference database for blast. We then
503 queried this via blastn with the SKn dataset to determine which of the 97 cHMTs were
504 either residents or very common transients of the nasal passages (Fig. 1, A). We
505 identified 30 cHMTs that were represented by ≥10 sequences in the SKn dataset with a
506 match at ≥98.5% identity. We added these 30 candidate taxa, represented by 31 16S
507 rRNA gene reference sequences for *e*HOMD (*e*HOMDrefs), as permanent HMTs to the
508 HOMDv14.51 alignment to generate the *e*HOMDv15.01 (Fig. 1, A and Table S7). Of
509 note, with the addition of nonoral taxa, we have replaced the old provisional taxonomy
510 prefix of Human Oral Taxon (HOT) with Human Microbial Taxon (HMT), which is applied
511 to all taxa in the *e*HOMD.
512 **Generating the provisional *e*HOMDv15.02 by identifying additional HMTs from a**
513 **dataset of 16S rRNA gene clones from human nostrils.** For the second step on the
514 HOMD expansion, we focused on obtaining new *e*HOMDrefs from the SKn dataset (i.e.,
515 the 44,374 16S rRNA gene clones from nostril (anterior nares) samples generated by
516 Julie Segre, Heidi Kong and colleagues (11-16)). We used blastn to query the SKn
517 clones versus the provisional database *e*HOMDv15.01. Of the nostril-derived 16S rRNA
518 gene clones, 37,716 of 44,374 matched reference sequences in *e*HOMDv15.01 at
519 ≥98.5% identity (Fig. 1, B) and 6163 matched to *e*HOMDv15.01 at <98% (Fig. 1, C).
520 The SKn clones that matched *e*HOMDv15.01 at ≥98.5% could be considered already
521 identified by *e*HOMDv15.01. Nevertheless, these already identified clones were used as
522 query to perform blastn versus the NCBI 16S database (27) to identify other NCBI
523 RefSeqs that might match these clones with a better identity. We compared the blastn

524 results against *e*HOMDv15.01 and NCBI 16S and if the match was substantially better
525 to a high-quality sequence (close to full length and without unresolved nucleotides) from
526 the NCBI 16S database then that one was considered for addition to the database.
527 Using this approach, we identified two new HMTs (represented by one *e*HOMDref each)
528 and five new *e*HOMDrefs for taxa present in *e*HOMDv14.51 that improved capture of
529 sequences to these taxa (Fig. 1, B and Table S7). For the 6163 SKn clones that
530 matched to *e*HOMDv15.01 at <98%, we performed clustering at ≥98.5% identity across
531 99% coverage and inferred an approximately maximum-likelihood phylogenetic tree
532 (Fig. 1, C and Supplemental Methods). If a cluster (an M-OTU) had ≥10 clone
533 sequences (30 out of 32), then we chose representative sequence(s) from that cluster
534 based on a visual assessment of the cluster alignment. Each representative sequence
535 was then queried against the NCBI nr/nt database to identify either the best high-quality,
536 named species-level match or, lacking this, the longest high-quality clone sequence to
537 use as the *e*HOMDref. Clones lacking a named match were assigned a genus name
538 based on their position in the tree and an HMT number, which serves as a provisional
539 name. The cluster representative sequence(s) plus any potentially superior reference
540 sequences from the NCBI nr/nt database were finally added to the *e*HOMDv15.01
541 alignment to create the *e*HOMDv15.02. Using this approach, we identified and added 28
542 new HMTs, represented in total by 38 *e*HOMDrefs (Fig. 1, C and Table S7). Of note, we
543 set aside the 1.1% (495 of 44,374) of SKn clones that matched at between 98 and
544 98.5% identify, to avoid calling a taxon where no new taxon existed in the tree-based
545 analysis of sequences that matched at <98%.

Generating the provisional *e*HOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. To further improve the performance of the evolving *e*HOMD, we took all of the SKn dataset clones that matched *e*HOMDv15.02 at <98.5% identity, clustered these at ≥98.5% identity across a coverage of 99% and inferred an approximately maximum-likelihood phylogenetic tree. Subsequent evaluation of this tree (see previous section) identified two more HMTs (represented in total by 3 *e*HOMDrefs) and one new *e*HOMDref for a taxon already in the database for addition to *e*HOMDv15.03 (Fig. 1, D and Table S7). To identify additional taxa that are resident to sites in the aerodigestive tract beyond the mouth and that are not represented by enough clones in the SKn dataset to meet our criteria, we iteratively evaluated the performance of *e*HOMDv15.02 with 5 other 16S rRNA gene datasets from aerodigestive tract sites outside the mouth (Fig. 1, E). We used the following criteria to select these datasets to assay for the performance of *e*HOMDv15.02 as a reference database for the aerodigestive tract across the span of human life in health and disease: (1) all sequences covered at least variable regions 1 and 2 (V1-V2), because for many bacteria resident in the aerodigestive tract V1-V2/V1-V3 includes sufficient sequence variability to get towards species-level assignment; and (2) the raw sequence data was either publicly available or readily supplied by the authors upon request. This approach yielded a representative set of datasets (Table S3) (21-23, 25, 26). Additional information on how we obtained and prepared each dataset for use is in Supplemental Methods. For each dataset from Table S3, we separately performed a blastn against *e*HOMDv15.02 and filtered the results to identify the percent of reads matching at ≥98.5% identity (Fig. 1, E). To compare the performance of

569 *e*HOMDv15.02 with other commonly used 16S rRNA gene databases, we also
570 performed a blastn against NCBI 16S (27), RDP16 (28) and SILVA128 (29, 30)
571 databases using the same filter as with *e*HOMDv15.02 for each dataset (Table S3). If
572 one of these other databases captured more sequences than *e*HOMDv15.02 at ≥98.5%
573 identity, we then identified the reference sequence in the outperforming database that
574 was capturing those sequences and evaluated it for inclusion in *e*HOMD. Based on this
575 comparative approach, we added three new HMTs (represented by one *e*HOMDref
576 each) plus five new eHOMDrefs for taxa already present in *e*HOMDv15.02 to the
577 provisional database to create *e*HOMDv15.03 (Fig. 1, E and Table S7).
578 **Generating the provisional *e*HOMDv15.04 by identifying additional candidate taxa**
579 **from a dataset of 16S rRNA gene clones from human skin.** Having established that
580 *e*HOMDv15.03 serves as an excellent 16S rRNA gene database for the aerodigestive
581 tract microbiome in health and disease, we were curious as to how it would perform
582 when evaluating 16S rRNA gene clone libraries from skin sites other than the nostrils.
583 As reviewed in (44), in humans, the area just inside the nostrils, which are the openings
584 into the nasal passages, is the skin-covered-surface of the nasal vestibule. Prior studies
585 have demonstrated that the bacterial microbiota of the skin of the nasal vestibule (a.k.a.
586 nostrils or nares) is distinctive and most similar to other moist skin sites (11). To test
587 how well *e*HOMDv15.03 performed as a database for skin microbiota in general, we
588 executed a blastn using 16S rRNA gene clones from all of the nonnasal skin sites
589 included in the Segre-Kong dataset (SKs) to assess the percentage of total sequences
590 captured at ≥98.5% identity over ≥98% coverage. Only 81.7% of the SKs clones were
591 identified with *e*HOMDv15.03, whereas 95% of the SKn clones were identified (Table

S2). We took the unidentified SKs sequences and did blastn versus the SILVA128 database with the same filtering criteria. To generate *e*HOMDv15.04, we first added the top 10 species from the SKs dataset that did not match to *e*HOMDv15.03, all of which had >350 reads in SKs (Fig. 1, F and Table S7). Of note, for two of the skin-covered body sites a single taxon accounted for the majority of reads that were unassigned with *e*HOMDv15.03: *Staphylococcus auricularis* from the external auditory canal and *Corynebacterium massiliense* from the umbilicus. Addition of these two considerably improved the performance of *e*HOMD for their respective body site. Next, we revisited the original list of 97 cHMTs and identified 4 species that are present in ≥3 of the 34 subjects in Kaspar et al. (19) (Table S1 column E), that had ≥30 reads in the SKs dataset and that matched to SILVA128 but not to *e*HOMDv15.03. These we added to generate *e*HOMDv15.04 (Fig. 1, G and Table S7).

Establishing *e*HOMD reference sequences and final updates to generate *e*HOMDv15.1. Each *e*HOMD reference sequence (*e*HOMDref) is a manually corrected representative sequence with a unique alphanumeric identifier that starts with its three-digit HMT #; each is associated with the original NCBI accession # of the candidate sequence. For each candidate 16S rRNA gene reference sequence selected, a blastn was performed against the NCBI nr/nt database and filtered for matches at ≥98.5% identity to identify additional sequences for comparison in an alignment, which was used to either manually correct the original candidate sequence or select a superior candidate from within the alignment. Manual correction included correction of all ambiguous nucleotides, any likely sequencing miscalls/errors and addition of consensus sequence at the 5'/3' ends to achieve uniform length. All ambiguous nucleotides from

615 earlier versions were corrected in the transition from HOMDv15.04 to *e*HOMDv15.1
616 because ambiguous bases, such as “R” and “Y”, are always counted as mismatches
617 against a nonambiguous base. Also, in preparing v15.1, nomenclature for
618 *Streptococcus* species was updated in accordance with (63) and genus names were
619 updated for species that were formerly part of the *Propionibacterium* genus in
620 accordance with (64). *Cutibacterium* is the new genus name for the formerly cutaneous
621 *Propionibacterium* species (64). In addition to the 79 taxa added in the expansion from
622 HOMDv14.51 to *e*HOMDv15.04 (Table S7), 4 oral taxa were added to the final
623 *e*HOMDv15.1: *Fusobacterium hwasookii* HMT-953, *Saccharibacteria* (TM7) bacterium
624 HMT-954, *Saccharibacteria* (TM7) bacterium HMT-955 and *Neisseria cinerea* HMT-956.
625 Also, *Neisseria pharyngis* HMT-729 was deleted because it is not validly named and is
626 part of the *N. sica–N. mucosa–N. flava* complex.
627 **Identification of taxa with a preference for the human nasal habitat**. We assigned
628 13 taxa as having the nostrils as their preferred body site habitat. To achieve this, we
629 first performed the following steps as illustrated in Table S6. 1) We performed blastn of
630 SKn and SKs versus *e*HOMDv15.04 and used the first hit to assign taxonomy to each
631 clone ; 2) used these names to generate a count table of taxa and body sites; 3)
632 normalized the total number of clones per body site to 20,000 each for comparisons
633 (columns B to V); 4) for each taxon, used the total number of clones across all body
634 sites as the denominator (column W) to calculate the % of that clone present at each
635 specific body site (columns Z to AT); 5) calculated the ratio of the % of each taxon in the
636 nostrils to the expected % if that taxon was evenly distributed across all 21 body sites in
637 the SKns clone dataset (column Y); and 6) sorted all taxa in Table S6 by the rank

638 abundance among the nostril clones (column X). Finally, of these top 20, we assigned
639 nasal as the preferred body site to those that were elevated ≥2x in the nostrils versus
640 what would be expected if evenly distributed across all the skin sites (column Y). This
641 conservative approach established a lower bound for the *e*HOMD taxa that have the
642 nasal passages as their preferred habitat. The SKn dataset includes samples from
643 children and adults in health and disease (11-16). In contrast, the HMP nares V1-V3
644 data are from adults 18 to 40 years of age in health only (23, 24). Of the species
645 classified as nasal in *e*HOMDv15.01, 8 of the 13 are in the top 19 most abundant
646 species from the 210-person HMP nares V1-V3 dataset.
647 **Reanalysis of the HMP nares V1-V3 dataset to species level.** We aligned the
648 2,338,563 chimera-cleaned reads present in the HMPnV1-V3 (see Suppl. Methods) in
649 QIIIME (pynast), using *e*HOMDv15.04 as reference database and trimmed for MED
650 using “o-trim-uninformative-columns-from-alignment” and “o-smart-trim” scripts (4).
651 2,203,471 reads (94.2% of starting) were recovered after the alignment and trimming
652 steps. After these initial cleaning steps, samples were selected such that only those with
653 more than 1000 reads were retained and each subject was represented by only one
654 sample. For subjects with more than one sample in the total HMP nares V1-V3 data, we
655 selected for use the one with more reads after the cleaning steps to avoid bias. Thus,
656 what we refer to as the HMP nares V1-V3 dataset included 1,627,514 high quality
657 sequences representing 210 subjects. We analyzed this dataset using MED with
658 minimum substantive abundance of an oligotype (-M) equal to 4 and maximum variation
659 allowed in each node (-V) equal to 12 nt, which equals 2.5% of the 820-nucleotide
660 length of the trimmed alignment. Of the 1,627,514 sequences, 89.9% (1,462,437)

661 passed the -M and -V filtering and are represented in the MED output. Oligotypes were
662 assigned taxonomy in R with the dada2::assignTaxonomy() function (an implementation
663 of the RDP naive Bayesian classifier algorithm with a kmer size of 8 and a bootstrap of
664 100) (5, 39) using the *e*HOMDv15.1 V1-V3 Training Set (version 1). We then collapsed
665 oligotypes within the same species/supraspecies yielding the data shown in Table S8.
666 The count data in Table S8 was converted to relative abundance by sample and relative
667 abundance was also calculated at each taxonomic level from Kingdom to
668 Species/Subgenus to generate an input table for linear discriminant analysis effect size
669 (LEfSe) (50). LEfSe (1.0.0-dev-e3cabe9) was performed using presence of *S. aureus*,
670 *Neisseriaceae* [G-1] bacterium HMT-174 or *L. clevelandensis* as class (-c) and per-
671 sample normalization (-o) of 1,000,000.
672 **Recruitment of genomes matching HMTs to *e*HOMD and assignment of species-**
673 **level names to genomes previously named only at then genus level.** Genomic
674 sequences were downloaded from the NCBI FTP site
675 (ftp://ftp.ncbi.nlm.nih.gov/genomes). Genome information, e.g., genus, species and
676 strain name were obtained from a summary file listed on the FTP site:
677 ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_genba
678 nk.txt. To recruit genomes for provisionally named *e*HOMD taxa (HMTs), genomic
679 sequences from the same genus were targeted. For 6 genera present in *e*HOMD, we
680 downloaded and analyzed 125 genomic sequences from GenBank that were
681 taxonomically assigned only to the genus level because some of these might belong to
682 a HMT. To determine the closest HMT for each of these genomes, the 16S rRNA genes
683 were extracted from each genome and were blastn-searched against the *e*HOMDv15.1

reference sequences. Of the 125 genomes tested, we excluded 13 that had <98% sequence identity to any of the *e*HOMDrefs. The remaining 112 genomes fell within a total of 27 *e*HOMD taxa at a percent identity ≥98.5% to one of the *e*HOMDrefs. To validate the phylogenetic relatedness of these genomes to HMTs, the extracted 16S rRNA gene sequences were then aligned with the *e*HOMDrefs using the MAFFT software (V6.935b) (65) and a phylogenetic tree was generated using FastTree (Version 2.1.9) (66) with the default Jukes-Cantor + CAT model for tree inference. The relationship of these genomes to *e*HOMD taxa was further confirmed by performing phylogenomic analysis in which all the proteins sequences of these genomes were collected and analyzed using PhyloPhlAn, which infers a phylogenomic tree based on the most conserved 400 bacterial protein sequences (38). These 112 genomes were then added to the *e*HOMDv15.1 as reference genomes. At least one genome from each taxon is dynamically annotated against a frequently updated NCBI nonredundant protein database so that potential functions may be assigned to hypothetical proteins due to matches to newly added proteins with functional annotation in NCBI nr database.

ACKNOWLEDGEMENTS

For supplying raw 16S rRNA gene tag sequences, we thank Melinda M. Pettigrew, Michele M. Sale and Emma Kaitlynn Allen. We are grateful to Vanja Klepac-Ceraj for thoughtful editing and commentary on the manuscript, to Hardeep Ranu for her help in keeping the project on pace, and to members of the Lemon Lab and the Starr-Dewhirst-Johnston-Lemon Joint Group Meeting for helpful questions and suggestions throughout the project.

Authorship contributions. Conceived Project: IFE, FED, KPL. Designed Project: IFE, TC, YH, FED, KPL. Analyzed data: IFE, YH, TC, FED, PG. Interpretted results: IFE, YH, TC, FED, KPL. Generated figures: IFE, TC. Wrote manuscript: KPL, IFE, FED, TC, YH. All authors approved the final manuscript.

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. The authors report no conflicts of interest.

REFERENCES

1. Bomar L, Brugger SD, Lemon KP. 2018. Bacterial microbiota of the nasal passages across the span of human life. Curr Opin Microbiol 41:8-14.
2. Conlan S, Kong HH, Segre JA. 2012. Species-level analysis of DNA sequence data from the NIH Human Microbiome Project. PLoS One 7:e47075.
3. Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. 2016. The Microbiome and the Respiratory Tract. Annu Rev Physiol 78:481-504.

730 4. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2015.
731 Minimum entropy decomposition: unsupervised oligotyping for sensitive
732 partitioning of high-throughput marker gene sequences. The ISME journal 9:968-
733 79.
734 5. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. 2016.
735 DADA2: High-resolution sample inference from Illumina amplicon data. Nat
736 Methods 13:581-3.
737 6. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The
738 Human Oral Microbiome Database: a web accessible resource for investigating
739 oral microbe taxonomic and genomic information. Database (Oxford)
740 2010:baq013.
741 7. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A,
742 Wade WG. 2010. The human oral microbiome. J Bacteriol 192:5002-17.
743 8. Eren AM, Borisy GG, Huse SM, Mark Welch JL. 2014. Oligotyping analysis of the
744 human oral microbiome. Proceedings of the National Academy of Sciences of the
745 United States of America 111:E2875-84.
746 9. Mark Welch JL, Utter DR, Rossetti BJ, Mark Welch DB, Eren AM, Borisy GG.
747 2014. Dynamics of tongue microbial communities with single-nucleotide
748 resolution using oligotyping. Front Microbiol 5:568.
749 10. Utter DR, Mark Welch JL, Borisy GG. 2016. Individuality, Stability, and Variability
750 of the Plaque Microbiome. Front Microbiol 7:564.
751 11. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG,
752 Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009.

753 Topographical and temporal diversity of the human skin microbiome. Science
754 324:1190-2.
755 12. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley
756 EC, Komarow HD, Murray PR, Turner ML, Segre JA. 2012. Temporal shifts in the
757 skin microbiome associated with disease flares and treatment in children with
758 atopic dermatitis. Genome research 22:850-9.
759 13. Oh J, Conlan S, Polley EC, Segre JA, Kong HH. 2012. Shifts in human skin and
760 nares microbiota of healthy children and adults. Genome medicine 4:77.
761 14. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos
762 E, Park M, Kong HH, Segre JA. 2013. Topographic diversity of fungal and
763 bacterial communities in human skin. Nature 498:367-70.
764 15. Oh J, Freeman AF, Park M, Sokolic R, Candotti F, Holland SM, Segre JA, Kong
765 HH. 2013. The altered landscape of the human skin microbiome in patients with
766 primary immunodeficiencies. Genome research 23:2103-14.
767 16. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. 2014. Biogeography
768 and individuality shape function in the human skin metagenome. Nature 514:59-
769 64.
770 17. Rasmussen TT, Kirkeby LP, Poulsen K, Reinholdt J, Kilian M. 2000. Resident
771 aerobic microbiota of the adult human nasal cavity. Apmis 108:663-75.
772 18. Boase S, Foreman A, Cleland E, Tan L, Melton-Kreft R, Pant H, Hu FZ, Ehrlich
773 GD, Wormald PJ. 2013. The microbiome of chronic rhinosinusitis: culture,
774 molecular diagnostics and biofilm detection. BMC Infect Dis 13:210.

19. Kaspar U, Kriegeskorte A, Schubert T, Peters G, Rudack C, Pieper DH, Wos-Oxley M, Becker K. 2016. The culturome of the human nose habitats reveals individual bacterial fingerprint patterns. Environ Microbiol 18:2130-42.

20. Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS. 2008. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 177:995-1001.

21. Laufer AS, Metlay JP, Gent JF, Fennie KP, Kong Y, Pettigrew MM. 2011. Microbial communities of the upper respiratory tract and otitis media in children. MBio 2:e00245-10.

22. Allen EK, Koeppel AF, Hendley JO, Turner SD, Winther B, Sale MM. 2014. Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge. Microbiome 2:22.

23. Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-14.

24. Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.

25. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ. 2004. Bacterial biota in the human distal esophagus. Proc Natl Acad Sci U S A 101:4250-5.

26. Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurso FJ, Pace NR. 2007. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. Proc Natl Acad Sci U S A 104:20529-33.

798 27. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,
799 Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y,
800 Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM,
801 Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W,
802 Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S,
803 Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H,
804 Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W,
805 Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at
806 NCBI: current status, taxonomic expansion, and functional annotation. Nucleic
807 Acids Res 44:D733-45.
808 28. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-
809 Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and
810 tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633-42.
811 29. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T,
812 Peplies J, Ludwig W, Glockner FO. 2014. The SILVA and "All-species Living
813 Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 42:D643-8.
814 30. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J,
815 Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved
816 data processing and web-based tools. Nucleic Acids Res 41:D590-6.
817 31. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P,
818 DeSantis TZ, Andersen GL, Wiener-Kronish JP, Bristow J. 2007. Loss of
819 bacterial diversity during antibiotic treatment of intubated patients colonized with
820 Pseudomonas aeruginosa. J Clin Microbiol 45:1954-62.

821 32. Perkins SD, Woeltje KF, Angenent LT. 2010. Endotracheal tube biofilm
822 inoculation of oral flora and subsequent colonization of opportunistic pathogens.
823 Int J Med Microbiol 300:503-11.
824 33. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW,
825 Carroll MP, Parkhill J, Bruce KD. 2011. Partitioning core and satellite taxa from
826 within cystic fibrosis lung bacterial communities. ISME J 5:780-91.
827 34. Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q,
828 Grice EA. 2016. Skin Microbiome Surveys Are Strongly Influenced by
829 Experimental Design. J Invest Dermatol 136:947-56.
830 35. Khamis A, Raoult D, La Scola B. 2004. *rpoB* gene sequencing for identification of
831 *Corynebacterium* species. J Clin Microbiol 42:3925-31.
832 36. Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis
833 of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.
834 Journal of microbiological methods 69:330-9.
835 37. Camarinha-Silva A, Jauregui R, Chaves-Moreno D, Oxley AP, Schaumburg F,
836 Becker K, Wos-Oxley ML, Pieper DH. 2014. Comparing the anterior nare
837 bacterial community of two discrete human populations using Illumina amplicon
838 sequencing. Environ Microbiol 16:2939-52.
839 38. Segata N, Bornigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new
840 method for improved phylogenetic and taxonomic placement of microbes. Nat
841 Commun 4:2304.

842 39. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for
843 rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl
844 Environ Microbiol 73:5261-7.
845 40. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG,
846 Angenent LT, Knight R, Ley RE. 2012. Impact of training sets on classification of
847 high-throughput bacterial 16s rRNA gene surveys. ISME J 6:94-103.
848 41. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A,
849 Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes
850 taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria
851 and archaea. ISME J 6:610-8.
852 42. Wade WG. 2015. Eubacterium, Bergey's Manual of Systematics of Archaea and
853 Bacteria doi:doi:10.1002/9781118960608.gbm00629.
854 43. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB,
855 Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B,
856 Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source,
857 platform-independent, community-supported software for describing and
858 comparing microbial communities. Appl Environ Microbiol 75:7537-41.
859 44. Brugger SD, Bomar L, Lemon KP. 2016. Commensal-Pathogen Interactions
860 along the Human Nasal Passages. PLoS pathogens 12:e1005633.
861 45. Akmatov MK, Koch N, Vital M, Ahrens W, Flesch-Janys D, Fricke J, Gatzemeier
862 A, Greiser H, Gunther K, Illig T, Kaaks R, Krone B, Kuhn A, Linseisen J,
863 Meisinger C, Michels K, Moebus S, Nieters A, Obi N, Schultze A, Six-Merker J,
864 Pieper DH, Pessler F. 2017. Determination of nasal and oropharyngeal

865 microbiomes in a multicenter population-based study - findings from Pretest 1 of
866 the German National Cohort. Sci Rep 7:1855.
867 46. Bell ME, Bernard KA, Harrington SM, Patel NB, Tucker TA, Metcalfe MG,
868 McQuiston JR. 2016. Lawsonella clevelandensis gen. nov., sp. nov., a new
869 member of the suborder Corynebacterineae isolated from human abscesses. Int
870 J Syst Evol Microbiol 66:2929-35.
871 47. Nicholson AC, Bell M, Humrighouse BW, McQuiston JR. 2015. Complete
872 Genome Sequences for Two Strains of a Novel Fastidious, Partially Acid-Fast,
873 Gram-Positive Corynebacterineae Bacterium, Derived from Human Clinical
874 Samples. Genome Announc 3.
875 48. Francuzik W, Franke K, Schumann RR, Heine G, Worm M. 2018.
876 Propionibacterium acnes Abundance Correlates Inversely with Staphylococcus
877 aureus: Data from Atopic Dermatitis Skin Microbiome. Acta Derm Venereol
878 98:490-495.
879 49. Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, Hu P,
880 Sodergren E, Liolios K, Huot-Creasy H, Birren BW, Earl AM. 2012. The "most
881 wanted" taxa from the human microbiome for whole genome sequencing. PLoS
882 One 7:e41294.
883 50. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower
884 C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol
885 12:R60.

886 51. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh
887 HA, Nouwen JL. 2005. The role of nasal carriage in *Staphylococcus aureus*
888 infections. Lancet Infect Dis 5:751-62.
889 52. Dantes R, Mu Y, Belflower R, Aragon D, Dumyati G, Harrison LH, Lessa FC,
890 Lynfield R, Nadle J, Petit S, Ray SM, Schaffner W, Townes J, Fridkin S,
891 Emerging Infections Program-Active Bacterial Core Surveillance MSI. 2013.
892 National burden of invasive methicillin-resistant Staphylococcus aureus
893 infections, United States, 2011. JAMA Intern Med 173:1970-8.
894 53. Gorwitz RJ, Kruszon-Moran D, McAllister SK, McQuillan G, McDougal LK,
895 Fosheim GE, Jensen BJ, Killgore G, Tenover FC, Kuehnert MJ. 2008. Changes
896 in the prevalence of nasal colonization with *Staphylococcus aureus* in the United
897 States, 2001-2004. J Infect Dis 197:1226-34.
898 54. Bode LG, Kluytmans JA, Wertheim HF, Bogaers D, Vandenbroucke-Grauls CM,
899 Roosendaal R, Troelstra A, Box AT, Voss A, van der Tweel I, van Belkum A,
900 Verbrugh HA, Vos MC. 2010. Preventing surgical-site infections in nasal carriers
901 of Staphylococcus aureus. N Engl J Med 362:9-17.
902 55. Proctor RA. 2015. Recent developments for Staphylococcus aureus vaccines:
903 clinical and basic science challenges. Eur Cell Mater 30:315-26.
904 56. Missiakas D, Schneewind O. 2016. Staphylococcus aureus vaccines: Deviating
905 from the carol. J Exp Med 213:1645-53.
906 57. Janek D, Zipperer A, Kulik A, Krismer B, Peschel A. 2016. High Frequency and
907 Diversity of Antimicrobial Activities Produced by Nasal Staphylococcus Strains
908 against Bacterial Competitors. PLoS Pathog 12:e1005812.

909 58. Zipperer A, Konnerth MC, Laux C, Berscheid A, Janek D, Weidenmaier C, Burian
910 M, Schilling NA, Slavetinsky C, Marschal M, Willmann M, Kalbacher H, Schittek
911 B, Brotz-Oesterhelt H, Grond S, Peschel A, Krismer B. 2016. Human
912 commensals producing a novel antibiotic impair pathogen colonization. Nature
913 535:511-6.
914 59. Nakatsuji T, Chen TH, Narala S, Chun KA, Two AM, Yun T, Shafiq F, Kotol PF,
915 Bouslimani A, Melnik AV, Latif H, Kim JN, Lockhart A, Artis K, David G, Taylor P,
916 Streib J, Dorrestein PC, Grier A, Gill SR, Zengler K, Hata TR, Leung DY, Gallo
917 RL. 2017. Antimicrobials from human skin commensal bacteria protect against
918 Staphylococcus aureus and are deficient in atopic dermatitis. Sci Transl Med 9.
919 60. Paharik AE, Parlet CP, Chung N, Todd DA, Rodriguez EI, Van Dyke MJ, Cech
920 NB, Horswill AR. 2017. Coagulase-Negative Staphylococcal Strain Prevents
921 Staphylococcus aureus Colonization and Skin Infection by Blocking Quorum
922 Sensing. Cell Host Microbe 22:746-756 e5.
923 61. Aguirre M, Morrison D, Cookson BD, Gay FW, Collins MD. 1993. Phenotypic and
924 phylogenetic characterization of some Gemella-like organisms from human
925 infections: description of Dolosigranulum pigrum gen. nov., sp. nov. J Appl
926 Bacteriol 75:608-12.
927 62. Liu CM, Price LB, Hungate BA, Abraham AG, Larsen LA, Christensen K, Stegger
928 M, Skov R, Andersen PS. 2015. Staphylococcus aureus and the ecology of the
929 nasal microbiome. Science Advances 1.
930 63. Jensen A, Scholz CF, Kilian M. 2016. Re-evaluation of the taxonomy of the Mitis
931 group of the genus Streptococcus based on whole genome phylogenetic

analyses, and proposed reclassification of Streptococcus dentisani as Streptococcus oralis subsp. dentisani comb. nov., Streptococcus tigurinus as Streptococcus oralis subsp. tigurinus comb. nov., and Streptococcus oligofermentans as a later synonym of Streptococcus cristatus. Int J Syst Evol Microbiol 66:4803-4820.

64. Scholz CF, Kilian M. 2016. The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. nov., Cutibacterium gen. nov. and Pseudopropionibacterium gen. nov. Int J Syst Evol Microbiol 66:4422-4432.

65. Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537:39-64.

66. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

FIGURE LEGENDS

Figure 1. The process for identifying Human Microbial Taxa (HMTs) from the aerodigestive tract to generate the *e*HOMD. Schematic of the approach used to identify taxa that were added as Human Microbial Taxa (HMT) to generate the *e*HOMDv15.04. Colored boxes are indicative of databases (blue), datasets (gray), newly added HMTs (green) and newly added *e*HOMDrefs for present HMTs (orange). Performance of blastn is indicated by yellow ovals and other tasks in white rectangles. HMT replaces the old HOMD taxonomy prefix HOT (human oral taxon). (**A**) Process for

generating the provisional *e*HOMDv15.01 by adding bacterial species from culture-dependent studies**.** (**B** and **C**) Process for generating the provisional *e*HOMDv15.02 by identifying additional HMTs from a dataset of 16S rRNA gene clones from human nostrils. (**D** and **E**) Process for generating the provisional *e*HOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. (**F** and **G**) Process for generating the provisional *e*HOMDv15.04 by identifying additional candidate taxa from a dataset of 16S rRNA gene clones from human skin. Please see Methods for detailed description of A–G. Abbreviations: NCBI 16S is the NCBI 16 Microbial database, *e*HOMDref is *e*HOMD reference sequence, db is database and ident is identity. Datasets included SKns (11-16), Allen et al. (22), Laufer et al. (21), Pei et al. (25) and Harris et al. (26). Kaspar et al. (19).

Figure 2. A small number of genera and species account for the majority of taxa in the HMP nares V1-V3 dataset at both an overall and individual level. Taxa identified in the reanalysis of the HMP nostril V1-V3 dataset graphed based on cumulative relative abundance of sequences at the genus- (**A**) and species/supraspecies- (**C**) level. The top 10 taxa are labeled. Prevalence (%) is indicated by the color gradient. The genus *Cutibacterium* includes species formerly known as the cutaneous *Propionibacterium* species, e.g., *P. acnes* (64). The minimum number of taxa at the genus- (**B**) and species/supraspecies- (**D**) level that accounted for 90% of the total sequences in each person's sample based on a table of taxa ranked by cumulative abundance from greatest to least. Ten or fewer species/supraspecies

977 accounted for 90% of the sequences in 94% of the 210 HMP participants in this
978 reanalysis.
979
980 **Figure 3. Common nasal species exhibit increased differential abundance when**
981 ***S. aureus* or *L. clevelandensis* are absent from the nostril microbiome.** In contrast,
982 no other species had differential abundance when *Neisseriaceae* [G-1] bacterium HMT-
983 174 was absent. Taxa at the genus- and species/supraspecies-level that showed
984 increased differential abundance in the reanalyzed HMP nares V1-V3 dataset when (**A**)
985 *S. aureus*, (**B**) *Neisseriaceae* [G-1] bacterium HMT-174 or (**C**) *L. clevelandensis* were
986 absent (dark gray bars) or present (light gray bars) based on LEfSe analysis (50). LDA
987 is Linear Discriminant Analysis and results with an LDA score of ≥ 3 are shown.
988
989 **TABLES**
990 **Table 1. The *e*HOMD outperforms comparable databases for species-level**
991 **taxonomic assignment to 16S rRNA reads from nostril samples (SKn dataset).**

Database	# Reads Identified[a]	% Reads Identified[a]
HOMDv14.5	22,274	50.2
*e*HOMDv15.1	42,197	**95.1**
NCBI 16S	38,337	86.4
RDP16	38,815	87.5
SILVA128	40,597	91.5

992 [a]Reads identified via blastn at 98.5% identity and 98% coverage

993 **Table 2. Performance of *e*HOMD and comparable databases for species-level taxonomic assignment to 16S rRNA gene**

994 **datasets from sites throughout the human aerodigestive tract.**

Dataset	16S Region	16S Primers	Sequecing Technique	Sample Type	# Samples	# Reads analyzed	Database	# Reads Identified[a]	% Reads Identified[a]
Laufer-Pettigrew (2011)	V1-V2	27F 338R	Roche/454	Nostril swab	108 children (108 samples)	120274	*e*HOMDv15.1	96233	80.0
							NCBI 16S	87082	72.4
							RDP16	97464	**81.0**
							SILVA128	97233	80.8
Allen-Sale (2014)	V1-V3	27F 534R	454-FLX	Nasal lavage fluid	10 adults (97 samples)	75310	*e*HOMDv15.1	68594	91.1
							NCBI 16S	63892	84.8
							RDP16	65028	86.4
							SILVA128	69082	**91.7**
Pei-Blaser (2004)	CL	318F 1519R 8F 1513R	CL	Esophageal biopsies	4 (10 libraries each)	7414	*e*HOMDv15.1	7276	**98.1**
							NCBI 16S	6686	90.2
							RDP16	6847	92.4
							SILVA128	7019	94.7
Harris-Pace (2007)	CL	27F 907R	CL	Brochial alveolar lavage fluid	57 children (50 libraries CF and 19 control)	3203	*e*HOMDv15.1	2684	**83.8**
							NCBI 16S	2427	75.8
							RDP16	2500	78.1
							SILVA128	2633	82.2
HMPnV1-V3	V1-V3	27F 534R	Roche/454	Nostril swab	227 adults (363 samples)[b]	2338563	*e*HOMDv15.1	2133083	**91.2**
							NCBI 16S	1932732	82.6
							RDP16	1965611	84.1
							SILVA128	2035882	87.1
vanderGast-Bruce (2011)	CL	7F 1510R	CL	Expectorated Sputa	14 adults (CF)	2137	*e*HOMDv15.1	2123	**99.3**
							NCBI 16S	2045	95.7
							RDP16	2057	96.3
							SILVA128	2084	97.5
Flanagan-Bristow (2007)	CL	27F 1492R	CL	Endotracheal tube aspirate	6 adults, 1 children (2-5 samples each)	3278	*e*HOMDv15.1	3193	97.4
							NCBI 16S	3186	97.2
							RDP16	3193	97.4
							SILVA128	3199	**97.6**

995

	Perkins-Angenent (2010)	CL	8F 1391R	CL	Extubated endotracheal tube	8 adults	1263	*e*HOMDv15.1	1008	**79.8**
								NCBI 16S	832	65.9
								RDP16	916	72.5
								SILVA128	1000	79.2

996 [a]Reads identified via blastn at 98.5% identity and 98% coverage

997 [b]See Supplemental Methods

998 CL = Clone library; CF = Cystic Fibrosis

Table 3. For nonnasal skin samples, the *e*HOMD performs best for species-level taxonomic assignment to 16S rRNA reads from oily skin sites (SKs dataset).

Skin_site	Skin type	Clones	*e*HOMD[a]	NCBI 16S[a]	RDP16[a]	SILVA128[a]	*e*HOMD minus SILVA
Alar crease	Oily	4149	**98.1**	82.1	82.3	95.4	**2.7**
External auditory canal	Oily	4970	**97.6**	87.4	87.6	90.2	**7.4**
Back	Oily	4552	**95.6**	92.2	92.2	92.5	**3.1**
Glabella	Oily	4287	**95.0**	79.8	80.4	92.5	**2.5**
Manubrium	Oily	4442	**93.5**	88.5	88.8	91.0	**2.5**
Retroauricular crease	Oily	15953	92.7	91.5	91.9	**93.4**	-0.7
Toe web space	Moist	4810	**89.4**	87.5	88.3	88.7	0.7
Occiput	Oily	8898	88.2	78.1	78.5	**88.4**	-0.2
Elbow	Dry	2181	**87.6**	76.5	77.1	78.1	**9.5**
Antecubital fossa	Moist	99077	85.4	85.4	86.9	**88.2**	-2.8
Gluteal crease	Moist	4656	**84.5**	81.5	83.2	84.3	0.2
Hypothenar palm	Dry	3650	84.5	89.1	87.9	**92.1**	-7.6
Inguinal crease	Moist	5031	**83.7**	82.3	81.6	83.5	0.2
Plantar heel	Moist	4013	82.8	82.4	83.2	**83.9**	-1.1
Volar forearm	Dry	92792	82.4	82.6	84.1	**85.7**	-3.3
Interdigital web space	Moist	3883	79.1	85.2	85.7	**88.3**	-9.2
Popliteal fossa	Moist	75284	78.5	83.8	84.9	**86.1**	-7.6
Buttock	Dry	4653	76.7	75.6	76.4	**77.6**	-0.9
Axillary vault	Moist	10148	72.1	70.7	72.3	**91.5**	-19.4
Umbilicus	Moist	4883	69.5	74.5	72.2	**76.4**	-6.9
TOTAL SKIN (Non-nasal sites: SKs)		362313	83.5	83.7	84.8	**86.7**	-3.2
Nostrils (nares; SKn)	Moist	44374	**95.1**	86.4	87.5	91.5	**3.6**

[a]Reads identified via blastn at 98.5% identity and 98% coverage

Color code: oily (blue), dry (red), and moist (green)

SUPPLEMENTAL MATERIALS

New insights into human nostril microbiome from the *expanded* Human Oral Microbiome Database (*e*HOMD): a resource for species-level identification of microbiome data from the aerodigestive tract

Isabel F. Escapa[a,b], Tsute Chen[a,b*], Yanmei Huang[a,b*], Prasad Gajare[a], Floyd E. Dewhirst[a,b], Katherine P. Lemon[a,c#]

Supplemental Figure Legends

Figure S1. Phylogenetic tree of *Betaproteobacteria* showing the positions of *Neisseriaceae* [G-1] bacterium HMT-174 and HMT-327. To see where the novel genus *Neisseriaceae* [G-1] fell relative other taxa at the family, class and order level, 10 non-oral sequences (in black font) were added to *e*HOMD sequences (in blue and red font) from the class *Betaproteobacteria* and a phylogenetic tree was generated. Species were selected from the families *Neisseriaceae* and *Chromobacteriaceae* (the two families in the order *Neisseriales*) because some of these sequences were best hits by simple blastn analysis of the novel *Neisseriaceae* [G-1] species. The tree was generated by first aligning the sequences with the MAFFT software (V6.935b) (1) and then subjecting them to FastTree (Version 2.1.9) (2) with the default Jukes-Cantor + CAT model for inferring the tree. The scale bar represents substitutions/site. Order names are marked above the appropriate node and *Neisseriales* families are indicated with brackets.

Supplemental Methods

Information on the aerodigestive tract microbiome datasets used.

Segre, Kong and colleagues have deposited close-to-full-length 16S RNA gene sequences from clone libraries collected from different skin sites, including the nostrils (nares) at NCBI under BioProjects PRJNA46333 and PRJNA30125 (3-8). We downloaded a total of 413,606 sequences from these BioProjects on May 11, 2017. The sequences were screened for bacterial 16S rRNA gene sequences only and parsed into two datasets: the SK nostril dataset (SKn), which includes 44,374 sequences from nostril samples with a mean length of 1354 bp (min. 1233, max. 1401); and the SK skin dataset (SKs), which includes 362,313 sequences with a mean length of 1356 bp (min. 1161, max. 1410). The SKs dataset includes 16S rRNA clone sequences derived from 20 non-nasal skin sites, including the alar crease, antecubital fossa, axillary vault, back, buttock, elbow, external auditory canal, glabella, gluteal crease, hypothenar palm, inguinal crease, interdigital web space, manubrium, occiput, plantar heel, popliteal fossa, retroauricular crease, toe web space, umbilicus and volar forearm.

The Human Microbiome Project (HMP) Data Coordination Center performed baseline processing and analysis of all 16S rRNA gene variable region sequences generated from >10,000 samples from healthy human subjects (9, 10). Table “HM16STR_healthy.csv” summarizes all the information for the 9811 files included in the dataset (https://www.hmpdacc.org/hmp/HM16STR/healthy/). We downloaded the 586 files labelled "anterior_nares" from the corresponding url identified in the same table. The downloaded files contain V1-V3, V3-V5 and V6-V9 data, therefore the reads

were filtered based on the primer information recorded on each read header, resulting in a total of 3,458,862 "anterior_nares" V1-V3 reads corresponding to 363 samples from 227 subjects. (See Methods for why the cohort used for species-level reanalysis included 210 subjects). We selected the 2,351,347 reads (67.9%) with length ≥430 and ≤652 bp (the range of the V1-V3 16S rRNA gene region in HOMDv14.51). After *de novo* chimera removal in QIIIME (USEARCH 6.1), there were 2,338,563 sequences for use. This dataset, dubbed HMPnV1-V3, was the starting point used to query the performance of the provisional versions of *e*HOMD and was the input for species-level reanalysis (see Methods).

Laufer et al. analyzed nostril swabs collected from 108 children ages 6 to 78 months in Philadelphia, PA between December 9, 2008 and January 2, 2009 for cultivation of *Streptococcus pneumoniae* and DNA harvest (11). Of these, 44% were culture positive for *S. pneumoniae* and 23% were diagnosed with otitis media. 16S rRNA gene V1-V2 sequences were generated using Roche/454 with primers 27F and 338R. We obtained 184,685 sequences from the authors, of which 94% included sequence matching primer 338R and 1% included sequence matching primer 27F. Therefore, we filtered reads for those ≥250 bp in length, quality score ≥30 and matching to 338R. We then performed demultiplexing in QIIME **(**split_libraries.py), eliminated sequences from samples for which there was no metadata (n=108 for metadata) leaving 120,963 sequences on which we performed *de novo* chimera removal in QIIME (USEARCH 6.1), yielding the 120,274 16S rRNA V1-V2 sequences used here.

Allen et al. collected nasal lavage fluid samples from 10 participants before, during and after experimental nasal inoculation with rhinovirus (12). 16S rRNA V1-V3 sequences

were generated using 454-FLX platform and primers 27F and 534R. We obtained 99,095 sequences from the authors of which 77,322 (78%) passed a length filter of ≥300 bp. After *de novo* chimera removal in QIIIME (USEARCH 6.1), there were 75,310 sequences for use in this study.

Pei et al. collected distal esophageal biopsies from four participants undergoing esophagogastroduodenoscopy for upper gastrointestinal complaints whose samples showed healthy esophageal tissue without evidence of pathology (13). From each of these, they generated ten 16s rRNA gene clone libraries from independent amplifications using two different primer pairs: 1) 318 to 1,519 with inosine at ambiguous positions and 2) from 8 to 1513. From this we downloaded 7,414 close-to-full-length 16S rRNA gene sequences from GenBank (GI: 109141477– 109141496).
Harris et al. collected bronchoalveolar lavage fluid from children with cystic fibrosis and generated 16S rRNA clone libraries from these (14). We downloaded these 3203 clones from GenBank (GI: AY805987–AY806453, DQ188268–DQ188805, and EU111806–EU112454).

van der Gast et al. generated 16S rRNA gene clone libraries from spontaneously expectorated sputum samples collected from 14 adults with cystic fibrosis (15). We downloaded these 2137 clones from GenBank (GI: FM995625–FM997761).

Flanagan et al. generated 16S rRNA gene clone libraries from daily endotracheal aspirates collected from seven intubated patients (16). We downloaded these 3278 clones from GenBank (GI: EF508731–EF512008).

Perkins et al. collected endotracheal tubes from eight adults with mechanical ventilation to generate 16S rRNA gene clone libraries (17). We downloaded these 1263 clones from GenBank (GI: FJ557249–FJ558511).

Information on the 16S rRNA gene databases used.

The NCBI 16S Microbial database (NCBI 16S) was downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/ on May 28, 2017 (18). RDP16 (rdp_species_assignment_16.fa.gz) and SILVA128 (silva_species_assignment_v128.fa.gz) files were downloaded from https://benjjneb.github.io/dada2/training.html and converted to BLAST databases using "makeblastdb" from the NCBI blast+ package (https://www.ncbi.nlm.nih.gov/books/NBK279690/) (19-21).

Addition of 16S rRNA sequences to the *e*HOMD alignment.

*e*HOMD maintains an alignment of all its reference 16S rRNA sequences. This alignment is based on the 16S rRNA secondary structure and is performed manually on a custom sequence editor (written in QuickBasic and available from FED). The corresponding alignment, in phylogenetic order, for each release of HOMD/*e*HOMD can be downloaded at http://www.homd.org/?name=seqDownload&type=R.

Clustering sequences at ≥98.5% and generating phylogenetic trees.

We performed blastn with an all-by-all search of the input sequences (Fig. 1, C and D). The blastn results were used to cluster the sequences into operational taxonomic units (OTUs) based on percent sequence identity and alignment coverage. Specifically, all

sequences were first sorted by size (seq_sort_len.fasta) in descending order and binned into operational taxonomic units (OTUs) at ≥98.5% identity across ≥99% coverage from longest to shortest sequences. If any subsequent sequence matched a previous sequence at ≥98.5% with coverage of ≥99%, the subsequent sequence was binned together with the previous sequence. If the subsequent sequence did not match any previous sequence, it was placed in new bin (i.e., 98.5% OTU). If the subsequent sequences matched multiple previous sequences that belong to more than one OTU, the subsequent sequence was binned to multiple OTUs, and at the same time, we formed a meta-OTU (M-OTU) linking these OTUs together. Next, we extracted sequences from each M-OTU and saved to individual fasta files. We then performed sequence alignment using software MAFFT (22) for each M-OTU fasta file and constructed phylogenetic trees for each M-OTU. The trees were built using FastTree (v2.1.9), which estimates nucleotide evolution with the Jukes-Cantor model and infers phylogenetic trees based on approximately maximum-likelihood (2). We organized the trees by using the longest branch as root and ordered from fewest nodes to more subnodes.

Additional information for cHMTs.

Of the 97 cHMTs for addition to HOMD, 82 are present in a nasal culturome of 34 participants (Table S1, column E), 18 with evidence of chronic nasal inflammation and 16 without evidence of nasal/systemic inflammation, based on swabs taken during nasal surgery from the anterior and posterior nasal vestibule (skin surface inside the nostrils) and the inferior and middle meatuses (23). Of the other 15 cHMTs we found 7 only in a report of cultivation of intraoperative mucosal swabs from 38 participants with chronic

rhinosinusitis (CRS) versus 6 controls (24); 7 only in sputa from 50 adults with CF (25); and 1 only in a report of the aerobic bacteria collected via a mucosal swab of the inferior turbinate and via a nasal wash from each of 10 healthy adults (26).

REFERENCES

1. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-66.
2. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.
3. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190-2.
4. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, Murray PR, Turner ML, Segre JA. 2012. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome research 22:850-9.
5. Oh J, Conlan S, Polley EC, Segre JA, Kong HH. 2012. Shifts in human skin and nares microbiota of healthy children and adults. Genome medicine 4:77.
6. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, Kong HH, Segre JA. 2013. Topographic diversity of fungal and bacterial communities in human skin. Nature 498:367-70.

7. Oh J, Freeman AF, Park M, Sokolic R, Candotti F, Holland SM, Segre JA, Kong HH. 2013. The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. Genome research 23:2103-14.

8. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature 514:59-64.

9. Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-14.

10. Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.

11. Laufer AS, Metlay JP, Gent JF, Fennie KP, Kong Y, Pettigrew MM. 2011. Microbial communities of the upper respiratory tract and otitis media in children. MBio 2:e00245-10.

12. Allen EK, Koeppel AF, Hendley JO, Turner SD, Winther B, Sale MM. 2014. Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge. Microbiome 2:22.

13. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ. 2004. Bacterial biota in the human distal esophagus. Proc Natl Acad Sci U S A 101:4250-5.

14. Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurso FJ, Pace NR. 2007. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. Proc Natl Acad Sci U S A 104:20529-33.

15. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW, Carroll MP, Parkhill J, Bruce KD. 2011. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. ISME J 5:780-91.

16. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P, DeSantis TZ, Andersen GL, Wiener-Kronish JP, Bristow J. 2007. Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with Pseudomonas aeruginosa. J Clin Microbiol 45:1954-62.

17. Perkins SD, Woeltje KF, Angenent LT. 2010. Endotracheal tube biofilm inoculation of oral flora and subsequent colonization of opportunistic pathogens. Int J Med Microbiol 300:503-11.

18. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-45.

19. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633-42.

20. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glockner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 42:D643-8.

21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-6.

22. Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537:39-64.

23. Kaspar U, Kriegeskorte A, Schubert T, Peters G, Rudack C, Pieper DH, Wos-Oxley M, Becker K. 2016. The culturome of the human nose habitats reveals individual bacterial fingerprint patterns. Environ Microbiol 18:2130-42.

24. Boase S, Foreman A, Cleland E, Tan L, Melton-Kreft R, Pant H, Hu FZ, Ehrlich GD, Wormald PJ. 2013. The microbiome of chronic rhinosinusitis: culture, molecular diagnostics and biofilm detection. BMC Infect Dis 13:210.

25. Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS. 2008. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 177:995-1001.

26. Rasmussen TT, Kirkeby LP, Poulsen K, Reinholdt J, Kilian M. 2000. Resident aerobic microbiota of the adult human nasal cavity. Apmis 108:663-75.

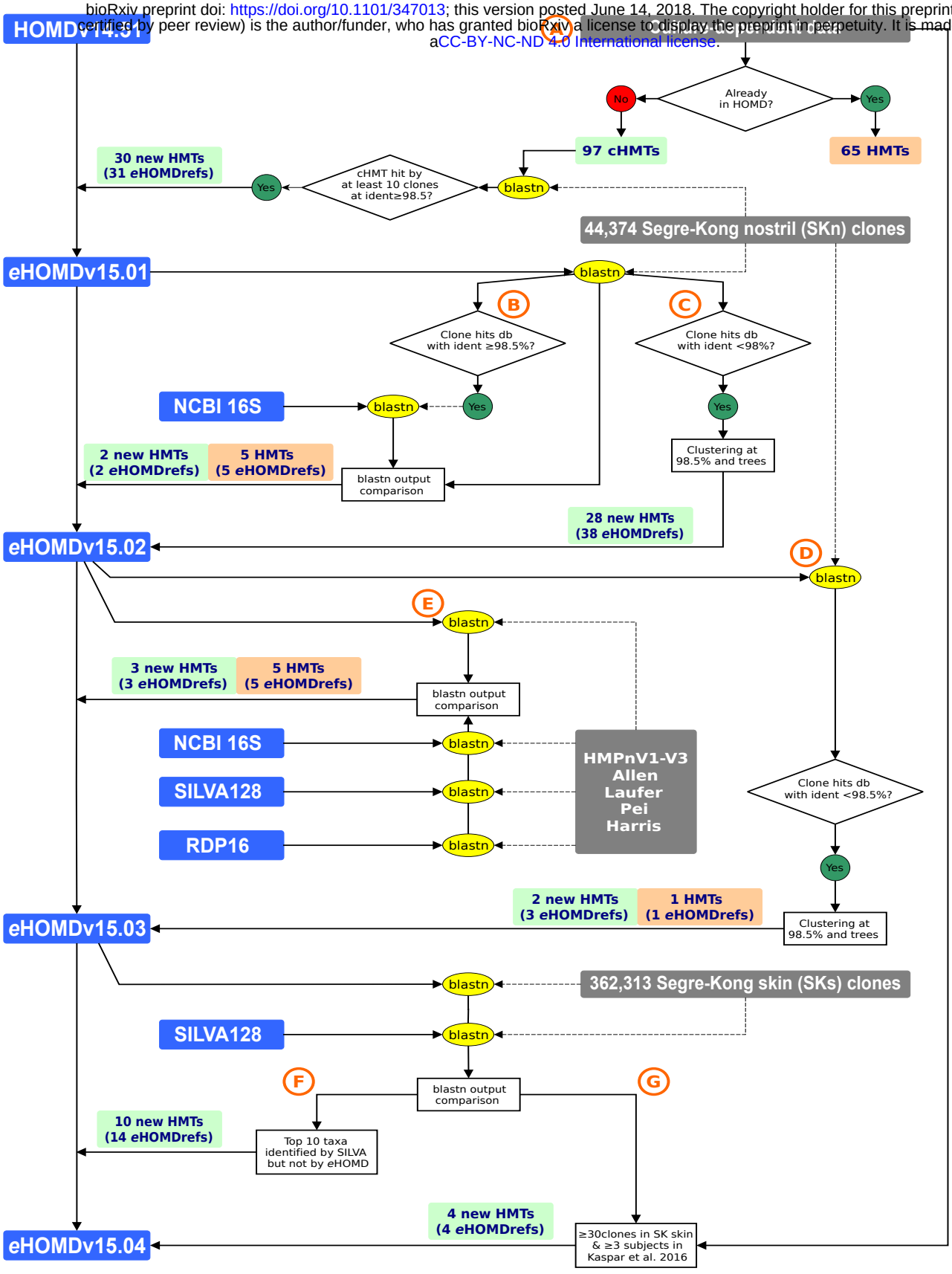

Figure 1. The process for identifying Human Microbial Taxa (HMTs) from the aerodigestive tract to generate the *e*HOMD. Schematic of the approach used to identify taxa that were added as Human Microbial Taxa (HMT) to generate the *e*HOMDv15.04. Colored boxes are indicative of databases (blue), datasets (gray), newly added HMTs (green) and newly added *e*HOMDrefs for present HMTs (orange). Performance of blastn is indicated by yellow ovals and other tasks in white rectangles. HMT replaces the old HOMD taxonomy prefix HOT (human oral taxon). (**A**) Process for generating the provisional *e*HOMDv15.01 by adding bacterial species from culture-dependent studies. (**B** and **C**) Process for generating the provisional *e*HOMDv15.02 by identifying additional HMTs from a dataset of 16S rRNA gene clones from human nostrils. (**D** and **E**) Process for generating the provisional *e*HOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. (**F** and **G**) Process for generating the provisional *e*HOMDv15.04 by identifying additional candidate taxa from a dataset of 16S rRNA gene clones from human skin. Please see Methods for detailed description of A–G. Abbreviations: NCBI 16S is the NCBI 16 Microbial database, *e*HOMDref is *e*HOMD reference sequence, db is database and ident is identity. Datasets included SKns (11-16), Allen et al. (22), Laufer et al. (21), Pei et al. (25) and Harris et al.

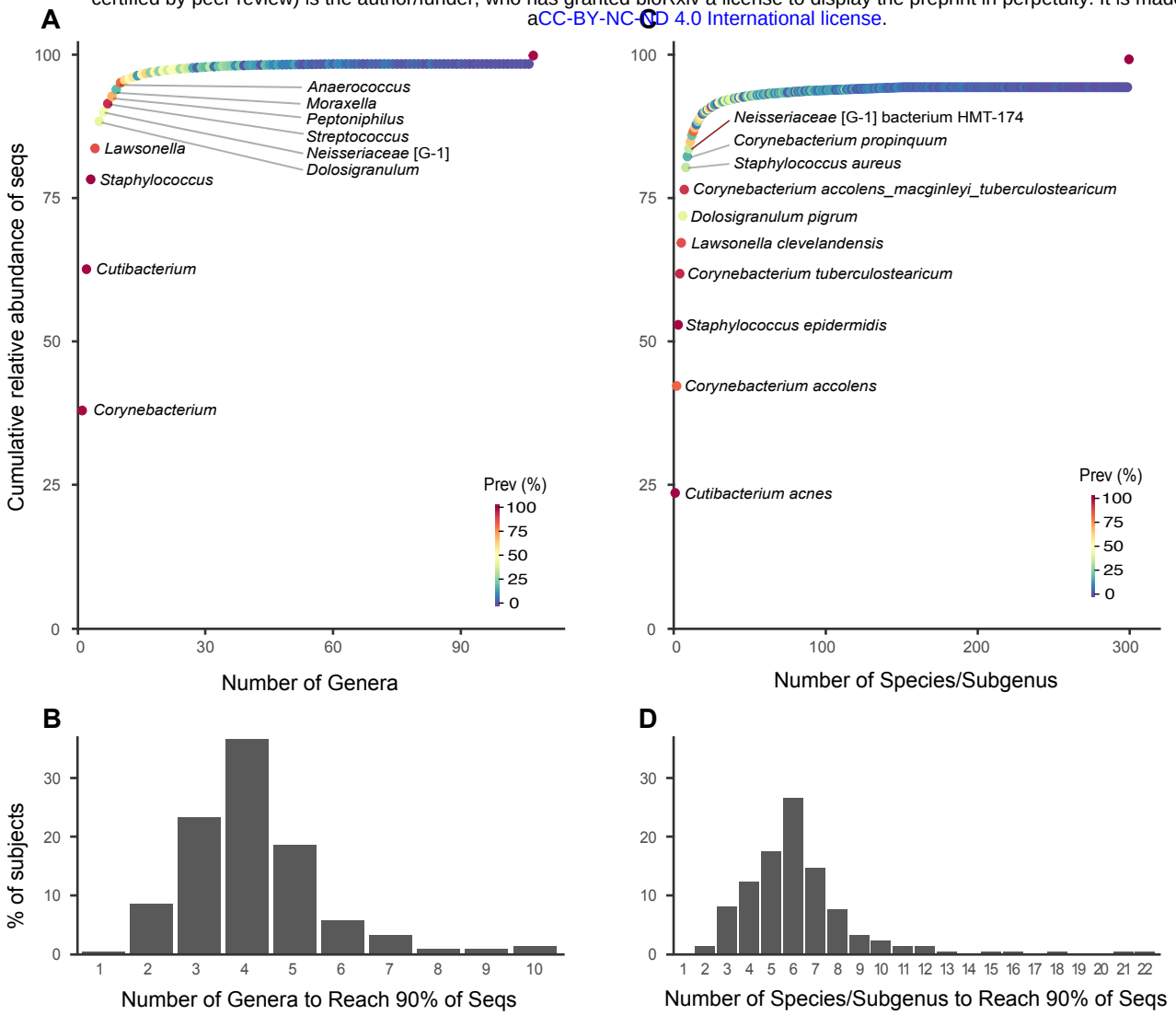

Figure 2. A small number of genera and species account for the majority of taxa in the HMP nares V1-V3 dataset at both an overall and individual level. Taxa identified in the reanalysis of the HMP nostril V1-V3 dataset graphed based on cumulative relative abundance of sequences at the genus- (**A**) and species/supraspecies- (**C**) level. The top 10 taxa are labeled. Prevalence (%) is indicated by the color gradient. The genus *Cutibacterium* includes species formerly known as the cutaneous *Propionibacterium* species, e.g., *P. acnes* (64). The minimum number of taxa at the genus- (**B**) and species/supraspecies- (**D**) level that accounted for 90% of the total sequences in each person's sample based on a table of taxa ranked by cumulative abundance from greatest to least. Ten or fewer species/supraspecies accounted for 90% of the sequences in 94% of the 210 HMP participants in this reanalysis.

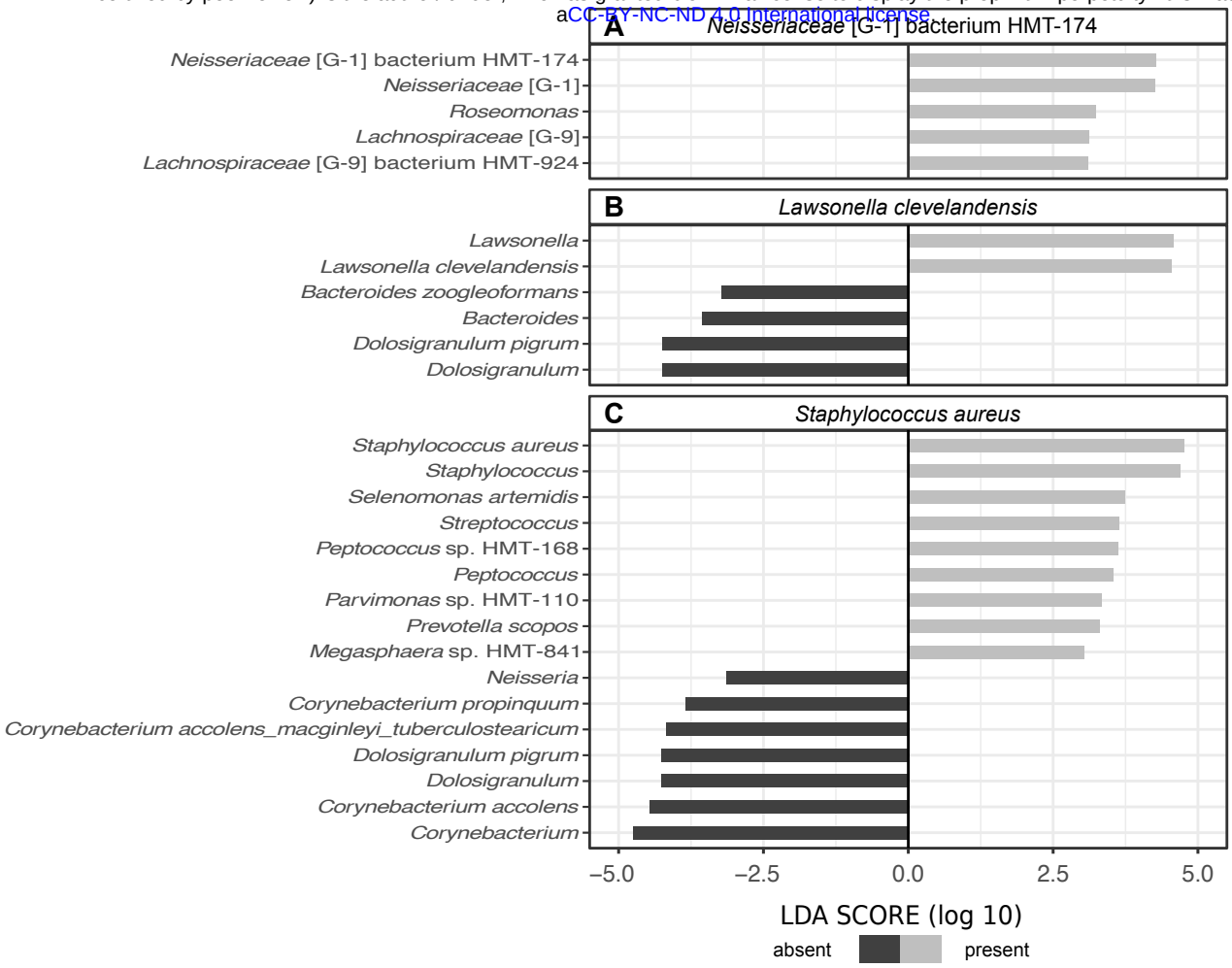

Figure 3. Common nasal species exhibit increased differential abundance when *S. aureus* or *L. clevelandensis* are absent from the nostril microbiome. In contrast, no other species had differential abundance when *Neisseriaceae* [G-1] bacterium HMT-174 was absent. Taxa at the genus- and species/supraspecies-level that showed increased differential abundance in the reanalyzed HMP nares V1-V3 dataset when (**A**) *S. aureus*, (**B**) *Neisseriaceae* [G-1] bacterium HMT-174 or (**C**) *L. clevelandensis* were absent (dark gray bars) or present (light gray bars) based on LEfSe analysis (50). LDA is Linear Discriminant Analysis and results with an LDA score of ≥ 3 are shown.