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

Isabel F. Escapa*<sup>a,b</sup>, Tsute Chen*<sup>a,b</sup>, Yanmei Huang*<sup>a,b</sup>, Prasad Gajare<sup>a</sup>, Floyd E. Dewhirst<sup>a,b</sup>, Katherine P. Lemon*<sup>a,c</sup>

<sup>a</sup>The Forsyth Institute (Microbiology), Cambridge, Massachusetts, USA
<sup>b</sup>Department of Oral Medicine, Infection & Immunity, Harvard School of Dental Medicine, Boston, Massachusetts, USA
<sup>c</sup>Division of Infectious Diseases, Boston Children’s Hospital, Harvard Medical School, Boston, Massachusetts, USA

Running Title: eHOMD: 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 (eHOMD) provides well-curated 16S rRNA gene reference sequences linked to available genomes. The eHOMD 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 eHOMD 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 eHOMD is a vital resource enhancing clinical relevance of 16S rRNA gene-based microbiome studies.

**IMPORTANCE**

The eHOMD ([ehomd.org](http://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 eHOMD is an actively curated, web-based, open-access resource. eHOMD 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 eHOMD 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 eHOMD becomes a comprehensive web-based resource for NGS microbiome analyses of body sites throughout the aerodigestive tract in both health and disease ([ehomd.org](http://ehomd.org)). The eHOMD 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 eHOMD 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 eHOMD 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 eHOMD, 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 eHOMDv15.1 (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 eHOMD.

RESULTS and DISCUSSION
I. The eHOMD is a Resource for Microbiome Research on the Human Upper Digestive and Respiratory Tracts.

As described below, the eHOMD 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 eHOMD facilitates the identification of aerodigestive and lower respiratory tract bacteria and provides phylogenetic, genomic, phenotypic, clinical and bibliographic information for these microbes.

The eHOMD outperforms other 16S rRNA gene databases for taxonomic identification of 16S rRNA gene clones from the human nostrils. Here we describe the generation of eHOMDv15.1, 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 eHOMDv15.01 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 eHOMDv15.02 (Fig. 1, B and C). Next, we evaluated how well eHOMDv15.02 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 eHOMDv15.02 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 eHOMDv15.03. 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 eHOMDv15.03 and SILVA128 (Fig. 1, F and G). Based on these results, we generated the eHOMDv15.1, 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 eHOMD with respect to
identification of clones from the SKn dataset, establishing eHOMD as a resource for the human nasal microbiome (Fig. 1 and Table S2).

The eHOMD is a resource for taxonomic assignment of 16S rRNA gene sequences from the entire human aerodigestive tract, as well as the lower respiratory tract. To assess the value of eHOMDv15.1 for analysis of datasets from sites throughout the human aerodigestive tract, it was compared with three commonly used 16S rRNA gene databases and consistently performed better than or comparable to these databases (Table 2). For these comparisons, we used blastn to assign taxonomy to three short-read (V1-V2 and V1-V3) and five approximately full-length-clone-library 16S rRNA gene datasets that are publicly available (21-23, 25, 26, 31-33). For short-read datasets, we focused on those covering all or part of the V1-V3 region of the 16S rRNA gene for three reasons. First, the V1-V3 region is preferentially used for the mouth. Second, for skin, V1-V3 sequencing results show high concordance with those from metagenomic sequencing (34). Third, to enable species-level distinctions within respiratory tract genera that include both common commensals and pathogens, V1-V3 is preferable for the nasal passages, sinuses and nasopharynx (2, 35-37). The chosen datasets include samples from children or adults in health and/or disease. The samples in these datasets are from human nostril swabs (21, 23), nasal lavage fluid (22), esophageal biopsies (25), extubated endotracheal tubes (32), endotracheal tube aspirates (31), sputa (33) and bronchoalveolar lavage (BAL) fluid (26). Endotracheal tube sampling may represent both upper and lower respiratory tract microbes and sputum may be contaminated by oral microbes, whereas BAL fluid represents microbes 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 eHOMD captures this variability. Similar to eHOMD, SILVA128 is also actively curated (29, 30). However, SILVA128 includes bacteria from a much larger range of habitats, whereas, eHOMD is focused on human-associated bacteria from the respiratory and upper digestive tracts. The performance of eHOMDv15.1 establishes it as a resource for microbiome studies of all body sites within the human respiratory and upper digestive tracts. The eHOMDv15.1 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 eHOMD might also perform well for other skin sites. To test this, we used eHOMDv15.04 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, eHOMDv15.04 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 eHOMDv15.04 consistently identified the most clones for each skin site to species level (98.5% identity and 98% coverage); eHOMDv15.04 is almost identical to the released eHOMDv15.1. In contrast, eHOMDv15.04 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
results revealed that a further expansion comparable to what we did here is necessary for eHOMD to include the full diversity of all skin sites.

The eHOMD is a resource for annotated genomes matched to HMTs for use in metagenomic and metatranscriptomic studies. Well-curated and annotated reference genomes correctly named at the species level are a critical resource for mapping metagenomic and metatranscriptomic data to gene and functional information, and for identifying species-level activity within the microbiome. There are currently >160,000 microbial genomic sequences deposited to GenBank; however, many of these genomes remain poorly or not-yet annotated or lack species-level taxonomy assignment, thus limiting the functional interpretation of metagenomic/metatranscriptomic studies to the genus level. Therefore, one goal of the eHOMD is to provide correctly named, curated and annotated genomes for all HMTs; this is an ongoing process. In generating eHOMDv15.1, we determined the species-level assignment for 112 genomes in GenBank that were previously identified only to the genus level and which matched to 27 eHOMD taxa. For each of these genomes, the phylogenetic relationship to the assigned HMT was verified by both phylogenetic analysis using 16S rRNA gene sequences and by phylogenomic analysis using a set of core proteins and PhyloPhIAn (38). To date, 85% (475) of the cultivated taxa (and 62% of all taxa), included in eHOMD have at least one sequenced genome.

The eHOMD is a resource for species-level assignment to the outputs of high-resolution 16S rRNA gene analysis algorithms. Newer algorithms, such as DADA2 and MED, permit high-resolution binning of 16S rRNA gene short-read sequences (4, 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 eHOMD, 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., \textit{Staphylococcus capitis\_caprae}.

**Advantages and limitations of the eHOMD.** The eHOMD 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 eHOMD 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 eHOMD. First, the eHOMD is based on extensively curated 16S rRNA reference sets (eHOMDrefs) 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 “\textit{Eubacteria}” in the phylum Firmicutes includes members that should be divided into multiple genera in 7 different families (42). In eHOMD, members of the “\textit{Eubacteria}” are placed in their phylogenetically appropriate family, e.g., \textit{Peptostreptococcaceae}, rather than incorrectly into the family \textit{Eubacteriaceae}. Appropriate taxonomy files are readily available from eHOMD for mothur (43) and other programs. Second, because eHOMD includes a provisional
species-level naming scheme, sequences that can only be assigned genus-level
taxonomy in other databases are resolved to species level via an HMT number. This
enhances the ability to identify and learn about taxa that currently lack full identification
and naming. Third, in eHOMD, for the 475 taxa with at least one sequenced genome,
genomes can be viewed graphically or searched using blastn, blastp or tblastn. For taxa
lacking available genomic sequences the available 16S rRNA sequences are included.
Many genomes of aerodigestive tract organisms are in the whole-genome shotgun
contigs (wgs) section of NCBI and are visible by blast search only through wgs provided
that you know the genome and can provide the BioProjectID or WGS Project ID. At
eHOMD, one can readily compare dozens to over a hundred genomes for some taxa to
begin to understand the pangenome of aerodigestive tract microbes. Fourth, we have
also compiled proteome sequence sets for genome-sequenced taxa enabling
proteomics and mass spectra searches on a dataset limited to proteins from ~2,000
relevant genomes.
In terms of limitations, the taxa included in the eHOMD, the 16S rRNA reference
sequences and genomes are not appropriate for samples from 1) human body sites
outside of the aerodigestive and respiratory tracts, 2) nonhuman hosts or 3) the
environment. In contrast, RDP (28), SILVA (29, 30) and Greengenes (41) are curated
16S rRNA databases inclusive of all sources and environments. Whereas, the NCBI
16S database is a curated set of sequences for bacterial and archaeal named species
only (a.k.a RefSeqs) that is more frequently updated (27). Finally, the NCBI nucleotide
database (nr/nt) includes the largest set of 16S rRNA sequences available; however,
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 eHOMD is preferred for the microbiomes of the human upper digestive and respiratory tracts.

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

**Identification of two previously unrecognized common nasal bacterial taxa.**

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

**The human nasal passages are the primary habitat for a subset of bacterial species.** The topologically external surfaces of the human body are the primary habitat for a number of bacterial taxa, which are often present at both high relative abundance
and high prevalence in the human microbiome. The previous HOMDv14.5 featured previously HOMDv14.5 featured putative body site habitat assignment for the majority of taxa included. In generating the eHOMDv15.1, we hypothesized that comparing the relative abundance of sequences identified to species or supraspecies level in the SKn clones and the SKs clones (nonnasal skin sites) would permit putative identification of the primary body-site habitat for a subset of nostril-associated bacteria. Based on criteria described in the methods, we identified 13 species as having the nostrils as their primary habitat (Table S6). Online at http://ehomd.org/index.php?name=HOMD the primary body site for each taxon is denoted as oral, nasal, skin, vaginal or unassigned. Definitive identification of the primary habitat of all human-associated bacteria will require species-level identification of bacteria at each distinct habitat across the surfaces of the human body from a cohort of individuals. This would enable a more complete version of the type of comparison performed here.

Members of the genus Corynebacterium (phylum Actinobacteria) are common in human nasal, skin and oral microbiomes but their species-level distribution across these body sites remains less clear (23). Our analysis of the SKns clones identified three Corynebacterium as primarily located in the nostrils compared to the other skin sites: C. propinquum, C. pseudodiphtheriticum and C. accolens (Table S6). In the species-level reanalysis of the HMP nares V1-V3 dataset, these were among the top five Corynebacterium species/supraspecies by rank order abundance of sequences (Table S5). In this reanalysis, Corynebacterium tuberculostearicum accounted for the fourth largest number of sequences; however, in the SKns clones it was not disproportionately present in the nostrils. Therefore, although common in the nostrils, we did not consider
the nostrils the primary habitat for *C. tuberculostearicium*, 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; Bell, 2016]. 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% O₂) 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 eHOMDv15.1, 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](http://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 eHOMD, 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
adult nostril microbiome supports future research to cultivate and characterize this bacterium. *Neisseriaceae [G-1] bacterium HMT-327* is another uncultivated nasal taxon, likely from the same unnamed genus, and the 20th (HMP) and 46th (SKn) most common nasal organism in the two datasets we reanalyzed. There are several additional uncultured nasal bacteria in eHOMD, highlighting the need for sophisticated cultivation studies even in the era of NGS studies. Having 16S rRNA reference sequences tied to the provisional taxonomic scheme in eHOMD allows targeted efforts to culture the previously uncultivated based on precise 16S rRNA identification methods.

**Species that are differentially abundant with respect to Neisseriaceae [G-1] bacterium HMT-174 or L. clevelandensis.** There is a lack of knowledge about potential relationships between the two newly recognized members of the nostril microbiome, *L. clevelandensis* and *Neisseriaceae [G-1] bacterium HMT-174*, and other known members of the nostril microbiome. Therefore, we performed LEfSe analysis on samples grouped based on the presence or absence of sequences of each of these two taxa of interest in search of species displaying differential relative abundance based on either of these two (50). For *Neisseriaceae [G-1] bacterium HMT-174*, this was targeted at identifying potential growth partners for this as-yet-uncultivated bacterium. In the presence of *Neisseriaceae [G-1] HMT-174, Lachnospiraceae [G-9] bacterium HMT-924* showed positive differential abundance (Fig. 3A). It had a low prevalence of 1.4% and low total relative abundance of 0.02% (Table S5). However, it also appears to be currently uncultivated. In contrast, a common nasal bacterium, *D. pigrum* showed positive differential abundance in the absence of *L. clevelandensis*, whereas no other 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 eHOMD 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
gene sequences corresponding to *S. aureus* were absent or present. Two of the five
most abundant genera in the adult nostrils, *Corynebacterium* and *Dolosigranulum* (Fig.
2A and Table S4), displayed the greatest positive differential abundance in the absence
of *S. aureus* (Fig. 3C dark gray bars). As previously reviewed (44), there is variability
between studies with smaller cohorts with respect to the reported correlations between
*S. aureus* and specific *Corynebacterium* species in the nostril microbiome; this
variability might relate to strain-level differences and/or to the small cohort sizes. In this
HMP cohort of 210 adults, 3 *Corynebacterium* species/supraspecies—*accolens*,
*accolens_macginleyi_tuberculostearicum* and *propinquum*—showed positive differential
abundance in the absence of *S. aureus* nostril colonization (Fig. 3C dark gray bars).
These 3 were among the 9 most abundant species in the cohort overall (Fig. 2C and
Table S5). Similarly, *D. pigrum* (61) showed a positive differential abundance in the
absence of *S. aureus* (Fig. 3C dark gray bars). This is consistent with observations from
Liu, Andersen and colleagues that high-levels of *D. pigrum* are the strongest predictor of
absence of *S. aureus* nostril colonization in 90 older adult Danish twin pairs (62). In our
reanalysis of the HMP nares V1-V3 dataset, *D. pigrum* was the 6th most abundant
species overall with a prevalence of 41% (Fig. 2C and Table S5). In contrast, the
species with positive differential abundance when *S. aureus* was present (Fig. 3C light
gray bars) all had a cumulative relative abundance of ≤0.01% and a prevalence of
≤3.33%, suggesting these are unlikely to be biologically relevant in most people
colonized by *S. aureus* (Table S5).

**Summary.** As demonstrated here, the eHOMD (ehomd.org) is a comprehensive well-
curated online database for the bacterial microbiome of the entire aerodigestive tract
enabling species/supraspecies-level taxonomic assignment to full-length and V1-V3 16S rRNA gene sequences and including correctly assigned, annotated available genomes. In generating the eHOMD, we identified two previously unrecognized common members of the adult human nostril microbiome, opening up new avenues for future research. As illustrated using the adult nostril microbiome, eHOMD can be leveraged for species-level analyses of the relationship between members of the aerodigestive tract microbiome, enhancing the clinical relevance of studies and generating new hypotheses about interspecies interactions and the functions of microbes within the human microbiome. The eHOMD provides a broad range of microbial researchers, from basic to clinical, a resource for exploring the microbial communities that inhabit the human respiratory and upper digestive tracts in health and disease.

MATERIALS AND METHODS

Generating the provisional eHOMDv15.01 by adding bacterial species from culture-dependent studies. To identify candidate Human Microbial Taxa (cHMTs), we reviewed two studies that included cultivation of swabs taken from along the nasal passages in both health and chronic rhinosinusitis (CRS) (18, 19) and one study of mucosal swabs and nasal washes only in health (17). We also reviewed a culture-dependent study of anaerobic bacteria isolated from cystic fibrosis (CF) sputa to identify anaerobes that might be present in the nasal passages/sinuses in CF (20). Using this approach, we identified 162 cHMTs, of which 65 were present in HOMDv14.51 and 97 were not (Fig. 1, A and Table S1). For each of these 97 named species, we downloaded
at least one 16S rRNA gene RefSeq from NCBI 16S (via a search of BioProjects 33175 and 33317) (27) and assembled these into a reference database for blast. We then queried this via blastn with the SKn dataset to determine which of the 97 cHMTs were either residents or very common transients of the nasal passages (Fig. 1, A). We identified 30 cHMTs that were represented by ≥10 sequences in the SKn dataset with a match at ≥98.5% identity. We added these 30 candidate taxa, represented by 31 16S rRNA gene reference sequences for eHOMD (eHOMDrefs), as permanent HMTs to the HOMDv14.51 alignment to generate the eHOMDv15.01 (Fig. 1, A and Table S7). Of note, with the addition of nonoral taxa, we have replaced the old provisional taxonomy prefix of Human Oral Taxon (HOT) with Human Microbial Taxon (HMT), which is applied to all taxa in the eHOMD.

Generating the provisional eHOMDv15.02 by identifying additional HMTs from a dataset of 16S rRNA gene clones from human nostrils. For the second step on the HOMD expansion, we focused on obtaining new eHOMDrefs from the SKn dataset (i.e., the 44,374 16S rRNA gene clones from nostril (anterior nares) samples generated by Julie Segre, Heidi Kong and colleagues (11-16)). We used blastn to query the SKn clones versus the provisional database eHOMDv15.01. Of the nostril-derived 16S rRNA gene clones, 37,716 of 44,374 matched reference sequences in eHOMDv15.01 at ≥98.5% identity (Fig. 1, B) and 6163 matched to eHOMDv15.01 at <98% (Fig. 1, C). The SKn clones that matched eHOMDv15.01 at ≥98.5% could be considered already identified by eHOMDv15.01. Nevertheless, these already identified clones were used as query to perform blastn versus the NCBI 16S database (27) to identify other NCBI RefSeqs that might match these clones with a better identity. We compared the blastn
results against eHOMDv15.01 and NCBI 16S and if the match was substantially better to a high-quality sequence (close to full length and without unresolved nucleotides) from the NCBI 16S database then that one was considered for addition to the database. Using this approach, we identified two new HMTs (represented by one eHOMDref each) and five new eHOMDrefs for taxa present in eHOMDv14.51 that improved capture of sequences to these taxa (Fig. 1, B and Table S7). For the 6163 SKn clones that matched to eHOMDv15.01 at <98%, we performed clustering at ≥98.5% identity across 99% coverage and inferred an approximately maximum-likelihood phylogenetic tree (Fig. 1, C and Supplemental Methods). If a cluster (an M-OTU) had ≥10 clone sequences (30 out of 32), then we chose representative sequence(s) from that cluster based on a visual assessment of the cluster alignment. Each representative sequence was then queried against the NCBI nr/nt database to identify either the best high-quality, named species-level match or, lacking this, the longest high-quality clone sequence to use as the eHOMDref. Clones lacking a named match were assigned a genus name based on their position in the tree and an HMT number, which serves as a provisional name. The cluster representative sequence(s) plus any potentially superior reference sequences from the NCBI nr/nt database were finally added to the eHOMDv15.01 alignment to create the eHOMDv15.02. Using this approach, we identified and added 28 new HMTs, represented in total by 38 eHOMDrefs (Fig. 1, C and Table S7). Of note, we set aside the 1.1% (495 of 44,374) of SKn clones that matched at between 98 and 98.5% identify, to avoid calling a taxon where no new taxon existed in the tree-based analysis of sequences that matched at <98%.
Generating the provisional eHOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. To further improve the performance of the evolving eHOMD, we took all of the SKn dataset clones that matched eHOMDv15.02 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 eHOMDrefs) and one new eHOMDref for a taxon already in the database for addition to eHOMDv15.03 (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 eHOMDv15.02 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 eHOMDv15.02 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 eHOMDv15.02 and filtered the results to identify the percent of reads matching at ≥98.5% identity (Fig. 1, E). To compare the performance of
eHOMDv15.02 with other commonly used 16S rRNA gene databases, we also performed a blastn against NCBI 16S (27), RDP16 (28) and SILVA128 (29, 30) databases using the same filter as with eHOMDv15.02 for each dataset (Table S3). If one of these other databases captured more sequences than eHOMDv15.02 at ≥98.5% identity, we then identified the reference sequence in the outperforming database that was capturing those sequences and evaluated it for inclusion in eHOMD. Based on this comparative approach, we added three new HMTs (represented by one eHOMDref each) plus five new eHOMDrefs for taxa already present in eHOMDv15.02 to the provisional database to create eHOMDv15.03 (Fig. 1, E and Table S7).

**Generating the provisional eHOMDv15.04 by identifying additional candidate taxa from a dataset of 16S rRNA gene clones from human skin.** Having established that eHOMDv15.03 serves as an excellent 16S rRNA gene database for the aerodigestive tract microbiome in health and disease, we were curious as to how it would perform when evaluating 16S rRNA gene clone libraries from skin sites other than the nostrils. As reviewed in (44), in humans, the area just inside the nostrils, which are the openings into the nasal passages, is the skin-covered-surface of the nasal vestibule. Prior studies have demonstrated that the bacterial microbiota of the skin of the nasal vestibule (a.k.a. nostrils or nares) is distinctive and most similar to other moist skin sites (11). To test how well eHOMDv15.03 performed as a database for skin microbiota in general, we executed a blastn using 16S rRNA gene clones from all of the nonnasal skin sites included in the Segre-Kong dataset (SKs) to assess the percentage of total sequences captured at ≥98.5% identity over ≥98% coverage. Only 81.7% of the SKs clones were identified with eHOMDv15.03, whereas 95% of the SKn clones were identified (Table
We took the unidentified SKs sequences and did blastn versus the SILVA128 database with the same filtering criteria. To generate eHOMDv15.04, we first added the top 10 species from the SKs dataset that did not match to eHOMDv15.03, 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 eHOMDv15.03: *Staphylococcus auricularis* from the external auditory canal and *Corynebacterium massiliense* from the umbilicus. Addition of these two considerably improved the performance of eHOMD 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 eHOMDv15.03. These we added to generate eHOMDv15.04 (Fig. 1, G and Table S7).

**Establishing eHOMD reference sequences and final updates to generate eHOMDv15.1.** Each eHOMD reference sequence (eHOMDref) 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
earlier versions were corrected in the transition from HOMDv15.04 to eHOMDv15.1 because ambiguous bases, such as “R” and “Y”, are always counted as mismatches against a nonambiguous base. Also, in preparing v15.1, nomenclature for \textit{Streptococcus} species was updated in accordance with (63) and genus names were updated for species that were formerly part of the \textit{Propionibacterium} genus in accordance with (64). \textit{Cutibacterium} is the new genus name for the formerly cutaneous \textit{Propionibacterium} species (64). In addition to the 79 taxa added in the expansion from HOMDv14.51 to eHOMDv15.04 (Table S7), 4 oral taxa were added to the final eHOMDv15.1: \textit{Fusobacterium hwasookii} HMT-953, \textit{Saccharibacteria} (TM7) bacterium HMT-954, \textit{Saccharibacteria} (TM7) bacterium HMT-955 and \textit{Neisseria cinerea} HMT-956. Also, \textit{Neisseria pharyngis} HMT-729 was deleted because it is not validly named and is part of the \textit{N. sica–N. mucosa–N. flava} complex.

\textbf{Identification of taxa with a preference for the human nasal habitat.} We assigned 13 taxa as having the nostrils as their preferred body site habitat. To achieve this, we first performed the following steps as illustrated in Table S6. 1) We performed blastn of SKn and SKs versus eHOMDv15.04 and used the first hit to assign taxonomy to each clone; 2) used these names to generate a count table of taxa and body sites; 3) normalized the total number of clones per body site to 20,000 each for comparisons (columns B to V); 4) for each taxon, used the total number of clones across all body sites as the denominator (column W) to calculate the \% of that clone present at each specific body site (columns Z to AT); 5) calculated the ratio of the \% of each taxon in the nostrils to the expected \% if that taxon was evenly distributed across all 21 body sites in the SKns clone dataset (column Y); and 6) sorted all taxa in Table S6 by the rank
abundance among the nostril clones (column X). Finally, of these top 20, we assigned nasal as the preferred body site to those that were elevated ≥2x in the nostrils versus what would be expected if evenly distributed across all the skin sites (column Y). This conservative approach established a lower bound for the eHOMD taxa that have the nasal passages as their preferred habitat. The SKn dataset includes samples from children and adults in health and disease (11-16). In contrast, the HMP nares V1-V3 data are from adults 18 to 40 years of age in health only (23, 24). Of the species classified as nasal in eHOMDv15.01, 8 of the 13 are in the top 19 most abundant species from the 210-person HMP nares V1-V3 dataset.

Reanalysis of the HMP nares V1-V3 dataset to species level. We aligned the 2,338,563 chimera-cleaned reads present in the HMPnV1-V3 (see Suppl. Methods) in QIIME (pynast), using eHOMDv15.04 as reference database and trimmed for MED using “o-trim-uninformative-columns-from-alignment” and “o-smart-trim” scripts (4). 2,203,471 reads (94.2% of starting) were recovered after the alignment and trimming steps. After these initial cleaning steps, samples were selected such that only those with more than 1000 reads were retained and each subject was represented by only one sample. For subjects with more than one sample in the total HMP nares V1-V3 data, we selected for use the one with more reads after the cleaning steps to avoid bias. Thus, what we refer to as the HMP nares V1-V3 dataset included 1,627,514 high quality sequences representing 210 subjects. We analyzed this dataset using MED with minimum substantive abundance of an oligotype (-M) equal to 4 and maximum variation allowed in each node (-V) equal to 12 nt, which equals 2.5% of the 820-nucleotide length of the trimmed alignment. Of the 1,627,514 sequences, 89.9% (1,462,437)
passed the -M and -V filtering and are represented in the MED output. Oligotypes were assigned taxonomy in R with the dada2::assignTaxonomy() function (an implementation of the RDP naive Bayesian classifier algorithm with a kmer size of 8 and a bootstrap of 100) (5, 39) using the eHOMDv15.1 V1-V3 Training Set (version 1). We then collapsed oligotypes within the same species/supraspecies yielding the data shown in Table S8. The count data in Table S8 was converted to relative abundance by sample and relative abundance was also calculated at each taxonomic level from Kingdom to Species/Subgenus to generate an input table for linear discriminant analysis effect size (LEfSe) (50). LEfSe (1.0.0-dev-e3cabe9) was performed using presence of S. aureus, Neisseriaceae [G-1] bacterium HMT-174 or L. clevelandensis as class (-c) and per-sample normalization (-o) of 1,000,000.

Recruitment of genomes matching HMTs to eHOMD and assignment of species-level names to genomes previously named only at then genus level. Genomic sequences were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes). Genome information, e.g., genus, species and strain name were obtained from a summary file listed on the FTP site: ftp://ftp.ncbi.nlm.nih.gov.genomes/ASSEMBLY_REPORTS/assembly_summary_genbank.txt. To recruit genomes for provisionally named eHOMD taxa (HMTs), genomic sequences from the same genus were targeted. For 6 genera present in eHOMD, we downloaded and analyzed 125 genomic sequences from GenBank that were taxonomically assigned only to the genus level because some of these might belong to a HMT. To determine the closest HMT for each of these genomes, the 16S rRNA genes were extracted from each genome and were blastn-searched against the eHOMDv15.1
reference sequences. Of the 125 genomes tested, we excluded 13 that had <98% sequence identity to any of the eHOMDrefs. The remaining 112 genomes fell within a total of 27 eHOMD taxa at a percent identity ≥98.5% to one of the eHOMDrefs. To validate the phylogenetic relatedness of these genomes to HMTs, the extracted 16S rRNA gene sequences were then aligned with the eHOMDrefs 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 eHOMD 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 eHOMDv15.1 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. Interpreted 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


microbiomes in a multicenter population-based study - findings from Pretest 1 of
the German National Cohort. Sci Rep 7:1855.

Bell ME, Bernard KA, Harrington SM, Patel NB, Tucker TA, Metcalfe MG, McQuiston JR. 2016. Lawsonella clevelandensis gen. nov., sp. nov., a new

Nicholson AC, Bell M, Humrighouse BW, McQuiston JR. 2015. Complete Genome Sequences for Two Strains of a Novel Fastidious, Partially Acid-Fast,
Gram-Positive Corynebacterineae Bacterium, Derived from Human Clinical Samples. Genome Announc 3.

aureus: Data from Atopic Dermatitis Skin Microbiome. Acta Derm Venereol 98:490-495.


63. Jensen A, Scholz CF, Kilian M. 2016. Re-evaluation of the taxonomy of the Mitis 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

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.

MAFFT. Methods Mol Biol 537:39-64.

66. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-

FIGURE LEGENDS

Figure 1. The process for identifying Human Microbial Taxa (HMTs) from the
aerodigestive tract to generate the eHOMD. Schematic of the approach used to
identify taxa that were added as Human Microbial Taxa (HMT) to generate the
eHOMDv15.04. Colored boxes are indicative of databases (blue), datasets (gray), newly
added HMTs (green) and newly added eHOMDrefs 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 eHOMDv15.01 by adding bacterial species from culture-dependent studies. (B and C) Process for generating the provisional eHOMDv15.02 by identifying additional HMTs from a dataset of 16S rRNA gene clones from human nostrils. (D and E) Process for generating the provisional eHOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. (F and G) Process for generating the provisional eHOMDv15.04 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, eHOMDref is eHOMD 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
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.

**TABLES**

**Table 1.** The eHOMD outperforms comparable databases for species-level taxonomic assignment to 16S rRNA reads from nostril samples (SKn dataset).

<table>
<thead>
<tr>
<th>Database</th>
<th># Reads Identified&lt;sup&gt;a&lt;/sup&gt;</th>
<th>% Reads Identified&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMDv14.5</td>
<td>22,274</td>
<td>50.2</td>
</tr>
<tr>
<td>eHOMDv15.1</td>
<td>42,197</td>
<td>95.1</td>
</tr>
<tr>
<td>NCBI 16S</td>
<td>38,337</td>
<td>86.4</td>
</tr>
<tr>
<td>RDP16</td>
<td>38,815</td>
<td>87.5</td>
</tr>
<tr>
<td>SILVA128</td>
<td>40,597</td>
<td>91.5</td>
</tr>
</tbody>
</table>

<sup>a</sup>Reads identified via blastn at 98.5% identity and 98% coverage
Table 2. Performance of eHOMD and comparable databases for species-level taxonomic assignment to 16S rRNA gene datasets from sites throughout the human aerodigestive tract.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>16S Region</th>
<th>16S Primers</th>
<th>Sequencing Technique</th>
<th>Sample Type</th>
<th># Samples</th>
<th># Reads analyzed</th>
<th>Database</th>
<th># Reads identifieda</th>
<th>% Reads identifieda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laufer-Pettigrew (2011)</td>
<td>V1-V2</td>
<td>27F 338R</td>
<td>Roche/454</td>
<td>Nostril swab</td>
<td>108 children (108 samples)</td>
<td>120274</td>
<td>eHOMDv15.1</td>
<td>96233</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>87082</td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>97464</td>
<td><strong>81.0</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>97233</td>
<td>80.8</td>
</tr>
<tr>
<td>Allen-Sale (2014)</td>
<td>V1-V3</td>
<td>27F 534R</td>
<td>454-FLX</td>
<td>Nasal lavage fluid</td>
<td>10 adults (97 samples)</td>
<td>75310</td>
<td>eHOMDv15.1</td>
<td>68594</td>
<td>91.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>63892</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>65028</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>69082</td>
<td><strong>91.7</strong></td>
</tr>
<tr>
<td>Pei-Blaser (2004)</td>
<td>CL</td>
<td>318F 1519R 8F 1513R</td>
<td>CL</td>
<td>Esophageal biopsies</td>
<td>4 (10 libraries each)</td>
<td>7414</td>
<td>eHOMDv15.1</td>
<td>7276</td>
<td><strong>98.1</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>6686</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>6847</td>
<td>92.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>7019</td>
<td>94.7</td>
</tr>
<tr>
<td>Harris-Pace (2007)</td>
<td>CL</td>
<td>27F 907R</td>
<td>CL</td>
<td>Brochial alveolar lavage fluid</td>
<td>57 children (50 libraries CF and 19 control)</td>
<td>3203</td>
<td>eHOMDv15.1</td>
<td>2684</td>
<td><strong>83.8</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>2427</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>2500</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>2633</td>
<td>82.2</td>
</tr>
<tr>
<td>HMPnV1-V3</td>
<td>V1-V3</td>
<td>27F 534R</td>
<td>Roche/454</td>
<td>Nostril swab</td>
<td>227 adults (363 samples)b</td>
<td>2338563</td>
<td>eHOMDv15.1</td>
<td>2133083</td>
<td><strong>91.2</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>1932732</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>1965611</td>
<td>84.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>2035882</td>
<td>87.1</td>
</tr>
<tr>
<td>vanderGast-Bruce (2011)</td>
<td>CL</td>
<td>7F 1510R</td>
<td>CL</td>
<td>Expectorated Sputa</td>
<td>14 adults (CF)</td>
<td>2137</td>
<td>eHOMDv15.1</td>
<td>2123</td>
<td><strong>99.3</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>2045</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>2057</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>2084</td>
<td>97.5</td>
</tr>
<tr>
<td>Flanagan-Bristow (2007)</td>
<td>CL</td>
<td>27F 1492R</td>
<td>CL</td>
<td>Endotracheal tube aspirate</td>
<td>6 adults, 1 children (2-5 samples each)</td>
<td>3278</td>
<td>eHOMDv15.1</td>
<td>3193</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCBI 16S</td>
<td>3186</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RDP16</td>
<td>3193</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SILVA128</td>
<td>3199</td>
<td><strong>97.6</strong></td>
</tr>
<tr>
<td>Study (2010)</td>
<td>Library</td>
<td>Forward Primer</td>
<td>Reverse Primer</td>
<td>Samples</td>
<td>Sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>----------------</td>
<td>----------------</td>
<td>---------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perkins-Angenent</td>
<td>CL 8F 1391R CL</td>
<td>Extubated endotracheal tube</td>
<td>8 adults</td>
<td>1263</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database</th>
<th>eHOMDv15.1</th>
<th>NCBI 16S</th>
<th>RDP16</th>
<th>SILVA128</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1008</td>
<td>832</td>
<td>916</td>
<td>1000</td>
</tr>
</tbody>
</table>

| Percentage | 79.8 | 65.9 | 72.5 | 79.2 |

*Reads identified via blastn at 98.5% identity and 98% coverage*

*See Supplemental Methods*

CL = Clone library; CF = Cystic Fibrosis
Table 3. For nonnasal skin samples, the eHOMD performs best for species-level taxonomic assignment to 16S rRNA reads from oily skin sites (SKs dataset).

<table>
<thead>
<tr>
<th>Skin_site</th>
<th>Skin type</th>
<th>Clones</th>
<th>eHOMD&lt;sup&gt;a&lt;/sup&gt;</th>
<th>NCBI 16S&lt;sup&gt;a&lt;/sup&gt;</th>
<th>RDP16&lt;sup&gt;a&lt;/sup&gt;</th>
<th>SILVA128&lt;sup&gt;a&lt;/sup&gt;</th>
<th>eHOMD minus SILVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alar crease</td>
<td>Oily</td>
<td>4149</td>
<td>98.1</td>
<td>82.1</td>
<td>82.3</td>
<td>95.4</td>
<td>2.7</td>
</tr>
<tr>
<td>External auditory canal</td>
<td>Oily</td>
<td>4970</td>
<td>97.6</td>
<td>87.4</td>
<td>87.6</td>
<td>90.2</td>
<td>7.4</td>
</tr>
<tr>
<td>Back</td>
<td>Oily</td>
<td>4552</td>
<td>95.6</td>
<td>92.2</td>
<td>92.2</td>
<td>92.5</td>
<td>3.1</td>
</tr>
<tr>
<td>Glabella</td>
<td>Oily</td>
<td>4287</td>
<td>95.0</td>
<td>79.8</td>
<td>80.4</td>
<td>92.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Manubrium</td>
<td>Oily</td>
<td>4442</td>
<td>93.5</td>
<td>88.5</td>
<td>88.8</td>
<td>91.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Retroauricular crease</td>
<td>Oily</td>
<td>15953</td>
<td>92.7</td>
<td>91.5</td>
<td>91.9</td>
<td>93.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>Toe web space</td>
<td>Moist</td>
<td>4810</td>
<td>89.4</td>
<td>87.5</td>
<td>88.3</td>
<td>88.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Occiput</td>
<td>Oily</td>
<td>8898</td>
<td>88.2</td>
<td>78.1</td>
<td>78.5</td>
<td>88.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>Elbow</td>
<td>Dry</td>
<td>2181</td>
<td>87.6</td>
<td>76.5</td>
<td>77.1</td>
<td>78.1</td>
<td>9.5</td>
</tr>
<tr>
<td>Antecubital fossa</td>
<td>Moist</td>
<td>99077</td>
<td>85.4</td>
<td>85.4</td>
<td>86.9</td>
<td>88.2</td>
<td>-2.8</td>
</tr>
<tr>
<td>Gluteal crease</td>
<td>Moist</td>
<td>4656</td>
<td>84.5</td>
<td>81.5</td>
<td>83.2</td>
<td>84.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Hypothenar palm</td>
<td>Dry</td>
<td>3650</td>
<td>84.5</td>
<td>89.1</td>
<td>87.9</td>
<td>92.1</td>
<td>-7.6</td>
</tr>
<tr>
<td>Inguinal crease</td>
<td>Moist</td>
<td>5031</td>
<td>83.7</td>
<td>82.3</td>
<td>81.6</td>
<td>83.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Plantar heel</td>
<td>Moist</td>
<td>4013</td>
<td>82.8</td>
<td>82.4</td>
<td>83.2</td>
<td>83.9</td>
<td>-1.1</td>
</tr>
<tr>
<td>Volar forearm</td>
<td>Dry</td>
<td>92792</td>
<td>82.4</td>
<td>82.6</td>
<td>84.1</td>
<td>85.7</td>
<td>-3.3</td>
</tr>
<tr>
<td>Interdigital web space</td>
<td>Moist</td>
<td>3883</td>
<td>79.1</td>
<td>85.2</td>
<td>85.7</td>
<td>88.3</td>
<td>-9.2</td>
</tr>
<tr>
<td>Popliteal fossa</td>
<td>Moist</td>
<td>75284</td>
<td>78.5</td>
<td>83.8</td>
<td>84.9</td>
<td>86.1</td>
<td>-7.6</td>
</tr>
<tr>
<td>Buttock</td>
<td>Dry</td>
<td>4653</td>
<td>76.7</td>
<td>75.6</td>
<td>76.4</td>
<td>77.6</td>
<td>-0.9</td>
</tr>
<tr>
<td>Axillary vault</td>
<td>Moist</td>
<td>10148</td>
<td>72.1</td>
<td>70.7</td>
<td>72.3</td>
<td>91.5</td>
<td>-19.4</td>
</tr>
<tr>
<td>Umbilicus</td>
<td>Moist</td>
<td>4883</td>
<td>69.5</td>
<td>74.5</td>
<td>72.2</td>
<td>76.4</td>
<td>-6.9</td>
</tr>
</tbody>
</table>

TOTAL SKIN (Non-nasal sites: SKs) | 362313 | 83.5 | 83.5 | 84.8 | 86.7 | -3.2 |

Nostrils (nares; SKn) | Moist | 44374 | 95.1 | 86.4 | 87.5 | 91.5 | 3.6 |

<sup>a</sup>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 (eHOMD): a resource for species-level identification of microbiome data from the aerodigestive tract

Isabel F. Escapa\textsuperscript{a,b}, Tsute Chen\textsuperscript{a,b,*}, Yanmei Huang\textsuperscript{a,b,*}, Prasad Gajare\textsuperscript{a}, Floyd E. Dewhirst\textsuperscript{a,b}, Katherine P. Lemon\textsuperscript{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, non-oral sequences (in black font) were added to eHOMD 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 eHOMD 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.**


**Addition of 16S rRNA sequences to the eHOMD alignment.**

eHOMD 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/eHOMD 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


Figure 1. The process for identifying Human Microbial Taxa (HMTs) from the aerodigestive tract to generate the eHOMD.

Schematic of the approach used to identify taxa that were added as Human Microbial Taxa (HMT) to generate the eHOMDv15.04. Colored boxes are indicative of databases (blue), datasets (gray), newly added HMTs (green) and newly added eHOMDrefs 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 eHOMDv15.01 by adding bacterial species from culture-dependent studies. (B and C) Process for generating the provisional eHOMDv15.02 by identifying additional HMTs from a dataset of 16S rRNA gene clones from human nostrils. (D and E) Process for generating the provisional eHOMDv15.03 by identifying additional candidate taxa from culture-independent studies of aerodigestive tract microbiomes. (F and G) Process for generating the provisional eHOMDv15.04 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, eHOMDref is eHOMD 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.